La struttura categoriale dell'intenzionalità

Funtori, trasformazioni naturali e i limiti del rappresentazionalismo

L'intenzionalità come struttura morfologica

I. Il problema dell'aboutness e le sue radici brentaniane

Franz Brentano, nelle *Psychologie vom empirischen Standpunkt* del 1874, formulò una tesi destinata a diventare il nodo gordiano attorno al quale si sarebbe avvolta tutta la filosofia della mente successiva: i fenomeni psichici si distinguono da quelli fisici per il fatto di essere intrinsecamente diretti verso un oggetto, di contenere in sé, per usare la locuzione scolastica che egli recuperò con consapevole anacronismo, una *inexistentia intentionale*. Ogni atto di credere, desiderare, percepire, immaginare è necessariamente un credere-che, un desiderare-qualcosa, un percepire-di. Questa struttura relazionale — che la tradizione analitica ha poi denominato semplicemente *aboutness* o intenzionalità — non è un accidente dei fenomeni mentali, ma la loro proprietà definitoria, ciò che li costituisce come mentali in senso proprio.

La ricezione analitica di questa idea è stata assai tormentata. Gottlob Frege, pur senza citare Brentano, introdusse una distinzione che avrebbe potuto risolvere almeno una parte del problema: quella tra *Sinn* e *Bedeutung*, tra il senso di un'espressione e il suo riferimento. Due espressioni possono avere lo stesso riferimento — il pianeta Venere — e nondimeno veicolare sensi diversi — «la stella del mattino» e «la stella della sera» —, tanto che la loro intercambiabilità non è garantita in tutti i contesti, e in particolare non nei contesti intenzionali: credere che la stella del mattino sia luminosa non equivale automaticamente a credere che la stella della sera lo sia, nemmeno quando le due espressioni denotano il medesimo oggetto astronomico. Il contenuto intenzionale, dunque, possiede una granularità che l'ontologia degli oggetti e delle proprietà non riesce da sola a catturare; esso determina come l'oggetto è presentato alla mente, e non soltanto quale oggetto è presentato.

Questa osservazione fregeana — che rimane una delle più solide acquisizioni della semantica filosofica novecentesca — apre però immediatamente un territorio concettuale di straordinaria complessità. Se il contenuto mentale si struttura attraverso

qualcosa di analogo al senso fregeano, che cosa è esattamente il senso? Quale tipo di entità popola lo spazio dei sensi, e in quale relazione sta con gli stati neurali del soggetto che li «ospita»? Come si organizzano i contenuti tra loro, e qual è la struttura dell'insieme dei contenuti possibili? A queste domande la semantica classica e la filosofia analitica della mente del XX secolo hanno proposto risposte molto diverse — dalle proposizioni di Russell alle proposizioni strutturate di Kaplan, dai modelli situazionali di Barwise e Perry alle teorie teleosemantiche di Millikan e Papineau —, ciascuna con i propri vantaggi e le proprie lacune caratteristiche.

L'ipotesi che si intende sviluppare in questo saggio è che esista uno strumento matematico — la teoria delle categorie, sviluppata da Samuel Eilenberg e Saunders Mac Lane a partire dal 1945 e da allora diventata il linguaggio unificante di larghe porzioni della matematica pura — che permette di riformulare il problema dell'intenzionalità a un livello di astrazione sufficiente da rendere visibili strutture che le teorie tradizionali lasciano nell'ombra. Non si tratta di ridurre la mente alla matematica, né di costruire un modello computazionale del cervello. Si tratta piuttosto di usare la grammatica categoriale come *lingua franca* che rivela le relazioni formali tra sistemi di contenuti, e poi di esaminare dove questa grammatica si inceppa — dove la struttura formale dell'intenzionalità eccede ciò che i morfismi sanno dire.

I.2 — Il labirinto delle teorie causali-informazionali

La teoria informazionale del contenuto mentale, nella sua versione più influente dovuta a Fred Dretske, sostiene che uno stato mentale M ha il contenuto C se e solo se M porta l'informazione che C è il caso, dove il concetto di informazione è inteso in senso shannoniano: M porta l'informazione che C nella misura in cui la probabilità condizionale $P(M \mid C)$ è sensibilmente più alta di $P(M)$. Questa formulazione possiede un'attrattiva immediata: assimila il significato mentale al significato naturale — le macchie di umidità significano pioggia, i cerchi negli alberi significano anni di vita — e permette di ancorare il contenuto a relazioni oggettive nel mondo fisico.

Il problema è noto ed è stato articolato da molti — Fodor, Cummins, Egan — in forme diverse. La versione più acuta riguarda la *disjunction problem*: se uno stato M è causato tanto dalla visione di una mucca quanto dalla visione di un cavallo nel buio che appare come una mucca, allora la teoria informazionale predice che il contenuto di M è *mucca-o-cavallo-mal-vista*, mentre l'intuizione robusta è che il contenuto sia semplicemente *mucca* e che il secondo caso costituisca un errore rappresentazionale. Dretske tenta di aggirare l'ostacolo distinguendo tra una fase di apprendimento — durante la quale lo stato acquisisce la sua funzione rappresentazionale — e una fase di dispiegamento, ma questa soluzione introduce una dipendenza dalla storia causale che rende il contenuto profondamente estrinseco e irriducibilmente relazionale in modi che risultano difficili da conciliare con la fenomenologia.

La risposta teleosemiotica di Ruth Millikan sposta il fuoco dall'informazione alla funzione biologica: uno stato M rappresenta C se la funzione propria di M — ciò per

cui *M* è stato selezionato nel corso dell'evoluzione — è produrre un comportamento adeguato in presenza di *C*. Questa strategia offre una soluzione elegante al problema della misrappresentazione (uno stato può avere il contenuto *C* senza che *C* sia presente, così come un cuore può avere la funzione di pompare sangue anche quando fallisce nel farlo), ma la paga in termini di un'opacità epistemica fondamentale: il contenuto di uno stato diventa una questione di storia evolutiva e di normatività biologica, dimensioni che risultano in linea di principio inaccessibili dall'interno della prospettiva del soggetto cognitivo.

Ancora più radicalmente, la teoria funzionalista nella sua versione computazionale — associata con i nomi di Hilary Putnam nella sua fase produttiva e di Jerry Fodor — identifica il contenuto con il ruolo funzionale di uno stato nell'economia complessiva del sistema cognitivo: il contenuto di *M* è determinato dalle relazioni inferenziali e causali che *M* intrattiene con gli input sensoriali, con gli output motori, e con gli altri stati interni. Si tratta di una proposta che ha il merito di rendere il contenuto compositivo e sistematico — esattamente come Fodor richiedeva — ma che paga questa virtù con la difficoltà di specificare quali relazioni funzionali siano rilevanti e con la nota critica dello «swampman» di Davidson: un duplicato funzionale di un essere umano, generato ex nihilo da un fulmine in una palude, sembrerebbe avere gli stessi contenuti mentali pur essendo privo di qualsiasi storia causale o funzionale.

Ciò che accomuna questi approcci, al di là delle loro profonde differenze, è il tentativo di ridurre il contenuto a una relazione: tra uno stato e un oggetto esterno (teorie causali), tra uno stato e una funzione biologica (teorie teleologiche), tra uno stato e altri stati (teorie funzionaliste). Ma le relazioni sono intrinsecamente *estrinseche:* esse non sono scritte nella natura interna dello stato, ma dipendono da come lo stato si colloca in una rete di connessioni che lo trascende. Questa estrinsecità genera inevitabilmente il problema del gap esplicativo nella sua forma più persistente: anche se si sapesse con precisione assoluta quali relazioni causali, funzionali o teleologiche uno stato intrattiene, sembrerebbe ancora concepibile che quello stato fosse privo di contenuto fenomenico, di quel *sapere che cos'è* avere quel contenuto di cui parla Thomas Nagel e che David Chalmers ha cristallizzato nella nozione di qualia.

I.3 — Diagnostica della difficoltà e orientamento verso una soluzione geometrica

La diagnosi che si propone è la seguente: le teorie standard del contenuto mentale falliscono perché trattano il contenuto come se fosse una proprietà *relazionale estrinseca* di stati fisici, quando invece vi sono buone ragioni — che si articoleranno nel corso di questo lavoro — per ritenere che il contenuto sia una proprietà *geometrica intrinseca* della struttura degli spazi in cui quegli stati vivono.

Questa affermazione richiede subito una precisazione cruciale, pena l'equivoco. Non si vuole sostenere che il contenuto sia riducibile alla geometria nel senso in cui la geometria è intesa nella fisica classica — come proprietà dello spazio-tempo o di qualche substrato fisico. La geometria di cui si parla è quella degli *spazi rappresentazionali*: le strutture matematiche che descrivono le relazioni di similarità,

distanza e curvatura tra le possibili configurazioni di uno stato cognitivo. Questi spazi sono *astratti* nel senso in cui lo sono le varietà riemanniane della geometria differenziale, ma sono anche *realizzati* nei sistemi cognitivi concreti nella misura in cui i pattern di attivazione neurale li istanziano.

L'intuizione guida — che si svilupperà con pieno rigore formale nella Parte Seconda e nella Parte Terza — può essere enunciata preliminarmente come segue. Consideriamo l'insieme di tutti i possibili stati di attivazione di un sistema neurale. Questo insieme non è un insieme amorfo: esso possiede una struttura ricca, che include relazioni di transizione (certi stati evolvono verso certi altri), relazioni di co-attivazione, relazioni di competizione e soppressione reciproca. Questa struttura conferisce all'insieme la forma di una *varietà differenziale* — un oggetto geometrico in cui ogni punto (ogni stato di attivazione) ha un intorno che assomiglia localmente a uno spazio euclideo di dimensione finita, ma la cui struttura globale può essere arbitrariamente complessa.

Ora, la tesi centrale è che il contenuto intenzionale di uno stato cognitivo è determinato non da relazioni estrinseche di quel stato con oggetti o processi esterni, bensì dalla struttura metrica e dalla curvatura della varietà rappresentazionale nel punto corrispondente a quello stato. Due stati cognitivi — realizzati in substrati fisici arbitrariamente diversi, magari in specie diverse o in architetture computazionali non biologiche — hanno lo stesso contenuto intenzionale se e solo se le varietà rappresentazionali in cui vivono sono *isometricamente equivalenti* nell'intorno rilevante. La *forma geometrica* dello spazio rappresentazionale *è* il contenuto.

Questa tesi ha almeno tre conseguenze filosoficamente rilevanti che si anticipano ora e si argomenteranno nei dettagli più avanti.

Prima conseguenza: il problema del content narrow versus wide — la disputa tra Putnam e Fodor circa se il contenuto sia determinato dall'interno del sistema cognitivo o dalle sue relazioni con l'ambiente — si dissolve parzialmente. La struttura geometrica intrinseca della varietà rappresentazionale è il contenuto *narrow*; le relazioni di embedding di quella varietà nell'ambiente fisico e sociale determinano una metrica *estrinseca* che corrisponde al contenuto *wide*. La distinzione è reale, ma diventa una distinzione geometrica ben definita — tra metrica intrinseca e metrica estrinseca di una varietà immersa — piuttosto che un dilemma filosofico senza soluzione tecnica.

Seconda conseguenza: la direzionalità dell'intenzionalità — il fatto che ogni atto mentale sia *su* qualcosa — trova una corrispondenza formale nel concetto di *flusso geodetico* sulla varietà rappresentazionale. L'oggetto intenzionale non è un'entità esterna che lo stato cognitivo cerca di agganciare tramite catene causali, ma l'*attrattore* del flusso geodetico generato dalla struttura metrica della varietà nello stato considerato. L'aboutness è, in questo senso, un fenomeno topologico prima ancora che semantico.

Terza conseguenza: il gap esplicativo tra descrizione fisica e contenuto fenomenico non scompare, ma si trasforma in un problema ben definito di *embedding*: il problema di capire in che misura la struttura metrica intrinseca di una varietà rappresentazionale possa essere ricavata dalla sua descrizione estrinseca in termini di attivazioni neurali misurabili. Questo problema ha una forma matematica precisa — è il problema

inverso dell'embedding di Whitney e le sue generalizzazioni — e può in linea di principio essere affrontato con strumenti tecnici concreti, a differenza del gap esplicativo nella sua formulazione standard.

Naturalmente, questa strategia apre anche nuove difficoltà. La più seria riguarda il cosiddetto problema dell'accesso: chi ha accesso alla struttura metrica intrinseca di una varietà rappresentazionale? Se solo il sistema cognitivo stesso ha accesso diretto a questa struttura — e i dati neuroscientifici suggeriscono che la misura delle attivazioni neurali ci dà al più una proiezione di dimensionalità ridotta della varietà reale — allora la teoria proposta rischia di essere verificabile solo dall'interno, il che la renderebbe una forma di idealismo trascendentale riformulato in linguaggio matematico. La risposta a questa difficoltà richiederà uno sviluppo tecnico che occuperà gran parte della Parte Quarta e della Parte Quinta. Per ora, si lasci questa preoccupazione sullo sfondo e si proceda alla costruzione dell'apparato formale necessario.

I.4 — Una nota metodologica sulla strategia di formalizzazione

Prima di addentrarsi nell'algebra e nella geometria, conviene chiarire la strategia metodologica. Non si propone una riduzione: non si sostiene che il contenuto mentale *sia* una struttura geometrica nello stesso senso in cui la temperatura di un gas *è* l'energia cinetica media delle molecole. Il rapporto è più sottile — è quello di una *rappresentazione formale* che, a differenza delle rappresentazioni proposizionali e relazionali dominanti nella letteratura, cattura aspetti strutturali del contenuto che altrimenti sfuggono.

L'analogia più utile è forse quella con la geometria dello spazio-tempo relativista. Prima di Einstein, il concetto di simultaneità sembrava ontologicamente semplice: due eventi sono simultanei o non lo sono. La teoria della relatività non ha eliminato la simultaneità, ma l'ha mostrata come relativa a un sistema di riferimento, e ha fornito gli strumenti geometrici — la metrica di Minkowski, le trasformazioni di Lorentz — per rendere conto in modo preciso di questo relativismo. Analogamente, si propone che il contenuto mentale non sia una proprietà assoluta degli stati cognitivi, ma una struttura geometrica relativa a una metrica rappresentazionale, e che la geometria dell'informazione fornisca gli strumenti per articolare questa struttura con precisione.

L'impresa non è senza rischi. La principale insidia è quella di produrre una formalizzazione nominale — di tradurre le intuizioni filosofiche in un linguaggio matematico senza che la matematica faccia un lavoro genuinamente esplicativo. Contro questo rischio ci si proteggerà richiedendo che la struttura formale sviluppata sia in grado di generare predizioni distinguibili da quelle delle teorie rivali e di suggerire esperimenti o simulazioni computazionali capaci in linea di principio di falsificarla. Questo standard di rigore è raramente incontrato nella filosofia della mente, ma è precisamente a questo standard che una proposta che si avvale di strumenti matematici ha il dovere di conformarsi.

II. Categorie, funtori e il linguaggio morfologico

Prima di procedere è necessario introdurre con precisione i concetti della teoria delle categorie che saranno impiegati. Una **categoria** $\mathscr{C}$ è costituita da due tipi di dati primitivi: una classe di *oggetti*, che si indica $\mathrm{Ob}(\mathscr{C})$, e per ogni coppia ordinata di oggetti (A, B) una classe di *morfismi* $\mathrm{Hom}_{\mathscr{C}}(A, B)$ — talvolta scritti $f: A \to B$ —, soggetti alle seguenti condizioni strutturali.

Per ogni oggetto A esiste un morfismo identità $\mathrm{id}_A \in \mathrm{Hom}(A, A)$. Per ogni coppia di morfismi $f: A \to B$ e $g: B \to C$ esiste un morfismo composto $g \circ f: A \to C$. La composizione è associativa:

$$h \circ (g \circ f) = (h \circ g) \circ f$$

e i morfismi identità sono neutrali rispetto alla composizione:

$$f \circ \mathrm{id}_A = f = \mathrm{id}_B \circ f \quad \text{per ogni } f: A \to B$$

Questo è tutto. La semplicità assiomatica è ingannevole, perché la potenza della teoria emerge non dalla ricchezza delle strutture interne agli oggetti — che possono restare del tutto opachi, come punti privi di struttura interna — bensì dalla rete dei morfismi che li connettono. Ciò che un oggetto *è*, nella prospettiva categoriale, si legge interamente in ciò che esso *fa*, ossia nei morfismi che lo coinvolgono come dominio e come codominio. Questa priorità dei morfismi sugli oggetti — che Mac Lane e Lawvere hanno elevato a vera e propria filosofia matematica — costituirà il filo conduttore dell'intera analisi.

Gli esempi paradigmatici di categorie nella pratica matematica includono: **Set**, la categoria degli insiemi con le funzioni come morfismi; **Top**, gli spazi topologici con le mappe continue; **Vect_k**, gli spazi vettoriali su un campo k con le mappe lineari; **Grp**, i gruppi con gli omomorfismi. Ma ciò che rende la teoria delle categorie filosoficamente interessante è la possibilità di costruire categorie i cui oggetti non siano strutture matematiche ma, per esempio, tipi linguistici, stati cognitivi, o situazioni percettive — categorie in cui i morfismi codificano relazioni di inferenza, implicazione, trasformazione prospettica o revisione doxastica.

La nozione successiva, cruciale per il nostro scopo, è quella di **funtore**. Dati due categorie $\mathscr{C}$ e $\mathscr{D}$, un funtore covariante $F: \mathscr{C} \to \mathscr{D}$ è un'applicazione che:

assegna a ogni oggetto $A \in \mathrm{Ob}(\mathscr{C})$ un oggetto $F(A) \in \mathrm{Ob}(\mathscr{D})$
assegna a ogni morfismo $f: A \to B$ in $\mathscr{C}$ un morfismo $F(f): F(A) \to F(B)$ in $\mathscr{D}$

preservando la struttura compositiva:

$F(g \circ f) = F(g) \circ F(f)$ per ogni $f: A \to B$, $g: B \to C$
$F(\mathrm{id}_A) = \mathrm{id}_{F(A)}$ per ogni $A \in \mathrm{Ob}(\mathscr{C})$

Un funtore, in altri termini, è una mappa che trasporta oggetti in oggetti e morfismi in morfismi rispettando le leggi della categoria sorgente nella categoria bersaglio. Esso non «distrugge» la struttura morfologica del dominio, ma la *proietta* nel codominio in modo coerente. Un funtore controvariante $F: \mathscr{C} \to \mathscr{D}$ procede analogamente, ma inverte la direzione dei morfismi:

$F(f): F(B) \to F(A)$ per ogni $f: A \to B$
$F(g \circ f) = F(f) \circ F(g)$

La terza nozione fondamentale è quella di **trasformazione naturale**. Se $F, G: \mathscr{C} \to \mathscr{D}$ sono due funtori dallo stesso dominio allo stesso codominio, una trasformazione naturale $\alpha: F \Rightarrow G$ è una famiglia di morfismi in $\mathscr{D}$, indicizzata sugli oggetti di $\mathscr{C}$:

$\{ \alpha_A: F(A) \to G(A) \}_{A \in Ob(\mathscr{C})}$

tale che per ogni morfismo $f: A \to B$ in $\mathscr{C}$ il seguente quadrato commuta:

```
F(A) ——F(f)——> F(B)
 |              |
α_A            α_B
 |              |
 v              v
G(A) ——G(f)——> G(B)
```

La condizione di naturalità — la commutatività di tutti questi quadrati — non è una richiesta arbitraria. Essa esprime l'idea che il passaggio da F a G sia *uniforme* rispetto alla struttura della categoria sorgente: non che F(A) venga trasformato in G(A) in modo capriccioso e dipendente da A, ma che la trasformazione si comporti coerentemente rispetto a tutti i morfismi del dominio. Storicamente, fu proprio la necessità di catturare questa nozione di «naturalità» a spingere Eilenberg e Mac Lane a inventare la teoria delle categorie: la nozione di isomorfismo naturale tra funtori era il loro obiettivo primario, e le categorie e i funtori furono introdotti come strumenti per dare ad essa una definizione precisa.

III. La categoria degli stati mentali: un'ontologia morfologica

Sia ora $\mathfrak{M}$ la categoria i cui oggetti sono *stati intenzionali* e i cui morfismi sono *transizioni inferenziali* tra di essi. Un oggetto di $\mathfrak{M}$ è, informalmente, un tipo di atto mentale — una credenza con un certo contenuto, un desiderio, una percezione, un'aspettativa — mentre un morfismo $f: M_1 \to M_2$ rappresenta una relazione di determinazione o implicazione cognitiva tra stati: M_2 è in qualche misura «accessibile» da M_1 attraverso un'operazione cognitiva determinata (inferenza deduttiva, aggiornamento bayesiano, associazione, memoria).

Questa costruzione è deliberatamente schematica, e occorre precisarla. La categoria $\mathfrak{M}$ non è intesa come un modello computazionale del cervello, né come una rappresentazione della psicologia folk. È piuttosto una struttura astratta che cattura la forma logica delle relazioni tra contenuti intenzionali, indipendentemente dal substrato che li realizza. In questo senso essa è più vicina alle strutture della semantica dei mondi possibili — che si possono intendere come categorie in cui gli oggetti sono mondi e i morfismi sono relazioni di accessibilità — che non a un modello neurobiologico.

Sia $\mathscr{W}$ la categoria dei «mondi o situazioni», i cui oggetti sono stati di cose — o configurazioni del mondo esterno — e i cui morfismi sono relazioni causali, mereologiche o di dipendenza nomologica tra di essi. L'intenzionalità, nella sua forma più elementare, è allora una struttura che connette queste due categorie. La proposta è di modellarla come un **funtore**:

$\Phi: \mathfrak{M} \to \mathscr{W}$

Un tale funtore assegna a ogni stato mentale M un suo contenuto intenzionale $\Phi(M)$ — uno stato di cose, un oggetto, una proprietà nel mondo — e a ogni transizione inferenziale $f: M_1 \to M_2$ una relazione $\Phi(f): \Phi(M_1) \to \Phi(M_2)$ tra i rispettivi contenuti. La condizione funtoriale impone che se M_2 segue inferenzialmente da M_1 per via di f, e M_3 segue da M_2 per via di g, allora il contenuto di M_3 deve essere raggiungibile dal contenuto di M_1 attraverso la composizione $\Phi(g) \circ \Phi(f)$. In altri termini: la struttura logica delle inferenze tra stati mentali si deve rispecchiare nella struttura delle relazioni tra i rispettivi contenuti mondani.

Questa è, in effetti, una versione formale della tesi della *normatività del contenuto* che Robert Brandom ha difeso in termini inferenziali: il contenuto di uno stato mentale è determinato dal ruolo che tale stato gioca nel sistema delle inferenze che il soggetto è disposto ad accettare. Se credo che tutti gli uomini siano mortali e credo che Socrate sia un uomo, il contenuto delle mie credenze è determinato anche dalla transizione inferenziale — il morfismo — che mi porta a concludere che Socrate è mortale.

Tuttavia il funtore Φ, da solo, non è sufficiente a rendere conto della distinzione fregeana tra senso e riferimento. Due funtori distinti $\Phi, \Psi: \mathfrak{M} \to \mathscr{W}$ possono coincidere su certi oggetti — assegnare lo stesso contenuto mondano a stati mentali diversi — e differire sui morfismi, rispecchiando il fatto che stati con lo stesso «riferimento» possono nondimeno avere strutture inferenziali diverse. La relazione tra Φ e Ψ, quando essi «puntano allo stesso mondo» su certi oggetti ma lo fanno attraverso percorsi morfologici distinti, è precisamente una **trasformazione naturale**:

$\alpha: \Phi \Rightarrow \Psi$

dove $\alpha_M: \Phi(M) \to \Psi(M)$ rappresenta, per ogni stato mentale M, la relazione tra il contenuto assegnatogli da Φ e quello assegnatogli da Ψ. La condizione di naturalità

garantisce che questo passaggio da una «prospettiva» intenzionale all'altra sia coerente con la struttura inferenziale di $\mathfrak{M}$.

La distinzione tra senso e riferimento riceve così una formulazione categoriale precisa: il **riferimento** di uno stato mentale è il valore dell'oggetto $\Phi(M) \in \mathscr{W}$, mentre il **senso** è, in senso tecnico, la *classe di equivalenza del funtore* rispetto alle trasformazioni naturali. Due funtori Φ e Ψ esprimono lo stesso senso se e solo se esiste una trasformazione naturale invertibile — un isomorfismo naturale — $\alpha: \Phi \Rightarrow \Psi$ tale che α_M sia un isomorfismo in $\mathscr{W}$ per ogni M. Essi possono riferirsi allo stesso oggetto (avere lo stesso «riferimento» su tutti gli oggetti di $\mathfrak{M}$) e nondimeno differire come funtori nel modo in cui strutturano morfologicamente le transizioni inferenziali.

Aggiunzioni, Yoneda e la soglia del fenomenico

IV. La struttura di aggiunzione e il ciclo percezione-azione

Tra i costrutti della teoria delle categorie, quello che più si avvicina a catturare la struttura del rapporto tra un sistema cognitivo e il suo ambiente è la nozione di **aggiunzione**. Date due categorie $\mathscr{C}$ e $\mathscr{D}$ e due funtori:

$L: \mathscr{D} \to \mathscr{C}$ (funtore sinistro)
$R: \mathscr{C} \to \mathscr{D}$ (funtore destro)

si dice che L è aggiunto sinistro di R — e che R è aggiunto destro di L, indicando il fatto con $L \dashv R$ — se esiste una biiezione naturale:

$\varphi\_{A,B}: \mathrm{Hom}\_{\mathscr{C}}(L(B), A) \cong \mathrm{Hom}\_{\mathscr{D}}(B, R(A))$

per ogni $A \in \mathrm{Ob}(\mathscr{C})$, $B \in \mathrm{Ob}(\mathscr{D})$, con la condizione che questa corrispondenza sia naturale in entrambi gli argomenti — ossia che per ogni $f: A \to A'$ in $\mathscr{C}$ e $g: B' \to B$ in $\mathscr{D}$ il diagramma rilevante commuti.

La naturalità dell'aggiunzione si può equivalentemente esprimere attraverso l'esistenza di due trasformazioni naturali — l'**unità** $\eta: \mathrm{Id}\_{\mathscr{D}} \Rightarrow R \circ L$ e il **counità** $\varepsilon: L \circ R \Rightarrow \mathrm{Id}\_{\mathscr{C}}$ — soggette alle cosiddette identità triangolari:

$(\varepsilon\_L) \circ (L\_\eta) = \mathrm{id}\_L$, ossia: $\varepsilon\_{L(B)} \circ L(\eta\_B) = \mathrm{id}\_{L(B)}$
$(R\_\varepsilon) \circ (\eta\_R) = \mathrm{id}\_R$, ossia: $R(\varepsilon\_A) \circ \eta\_{R(A)} = \mathrm{id}\_{R(A)}$

Queste identità esprimono una simmetria profonda: che il «viaggio» da $\mathscr{D}$ a $\mathscr{C}$ via L, seguito dal rientro in $\mathscr{D}$ via R, e poi ancora in $\mathscr{C}$ via ε, dà luogo a un percorso che si annulla nella struttura dell'identità. L'aggiunzione è, in un senso tecnico preciso, la forma più debole di «inversa reciproca» tra funtori — meno forte dell'equivalenza di

categorie (che richiederebbe isomorfismi naturali nelle identità triangolari invece di semplici morfismi), ma abbastanza robusta da generare strutture matematiche di grandissima ricchezza.

La proposta interpretativa è la seguente: il ciclo percezione-azione che costituisce il rapporto fondamentale di un organismo con il suo ambiente possiede la struttura formale di un'aggiunzione. Sia $\mathscr{W}$ la categoria degli stati del mondo, sia $\mathfrak{M}$ quella degli stati mentali come definita in precedenza. Il processo percettivo è un funtore:

$$P: \mathscr{W} \to \mathfrak{M}$$

che trasporta stati del mondo in rappresentazioni mentali: P(w) è la rappresentazione che l'organismo forma dello stato w attraverso il proprio apparato sensoriale. Il processo attivo — l'azione — è invece un funtore in direzione opposta:

$$A: \mathfrak{M} \to \mathscr{W}$$

che trasporta stati mentali in stati del mondo: A(m) è lo stato del mondo che l'organismo tende a produrre quando si trova nello stato mentale m.

L'affermazione che $A \dashv P$ — che A è aggiunto sinistro di P — significa che esiste una corrispondenza naturale:

$$\mathrm{Hom}_{\mathscr{W}}(A(m), w) \cong \mathrm{Hom}_{\mathfrak{M}}(m, P(w))$$

Questa biiezione ha un'interpretazione cognitiva precisa: un'azione che produca lo stato del mondo w a partire dallo stato mentale m corrisponde biunivocamente a un'inferenza che, a partire da m, porti alla rappresentazione P(w) di quel medesimo stato. Il fatto che lo spazio dei morfismi sia il medesimo dei due lati dell'aggiunzione — che il «prezzo» di passare da m a w attraverso l'azione sia identico al «prezzo» di passare da m alla sua rappresentazione P(w) attraverso l'inferenza — cattura qualcosa di fondamentale nella struttura del comportamento adattivo: l'organismo agisce sul mondo esattamente nella misura in cui la propria rappresentazione del mondo lo «spinge» verso determinate trasformazioni. Non c'è, a questo livello di analisi, una separazione netta tra il rappresentare e il fare.

L'unità $\eta: \mathrm{Id}_{\mathfrak{M}} \Rightarrow P \circ A$ esprime allora il fatto che ogni stato mentale è mediato dalla percezione degli effetti che esso stesso tende a generare: l'organismo non ha accesso diretto al mondo, ma soltanto alla propria rappresentazione percettiva degli stati che le proprie azioni inducono — una tesi che risuona potentemente con la concezione helmholtziana della percezione come inferenza inconscia, e con il suo erede formale nelle teorie del cervello predittivo. Il counità $\varepsilon: A \circ P \Rightarrow \mathrm{Id}_{\mathscr{W}}$ esprime invece il fatto che l'azione tende a realizzare nel mondo ciò che la percezione rappresenta: gli stati del mondo raggiungibili attraverso il ciclo $A \circ P$ — agire sulla base di ciò che si percepisce

— si avvicinano agli stati del mondo effettivi nella misura in cui il sistema è ben adattato al proprio nicho ecologico.

La struttura di aggiunzione non è una metafora. È un vincolo matematico preciso che, se soddisfatto, implica una serie di proprietà strutturali — la preservazione di certi limiti e colimiti, l'esistenza di soluzioni universali a problemi di ottimizzazione — che sono esattamente le proprietà che ci si aspetta da un sistema cognitivo adattivo. La «ottimalità» biologica, in questa lettura, è la traccia categoriale dell'aggiunzione.

V. Il lemma di Yoneda e l'identità attraverso le relazioni

Il risultato più profondo e, ai fini della nostra indagine, più rilevante della teoria delle categorie elementare è il **lemma di Yoneda**. Prima di enunciarlo occorre introdurre la nozione di **categoria dei funtori** e quella di **funtore rappresentabile**.

Data una piccola categoria $\mathscr{C}$ (per semplicità, una categoria la cui classe di oggetti è un insieme), si può formare la categoria Fun($\mathscr{C}$^{op}, Set) — talvolta indicata come la **categoria dei prefasci** su $\mathscr{C}$, o con la notazione $\hat{\mathscr{C}}$ — i cui oggetti sono i funtori controvarianti da $\mathscr{C}$ a Set e i cui morfismi sono le trasformazioni naturali tra tali funtori.

Per ogni oggetto $A \in \mathrm{Ob}(\mathscr{C})$ si definisce il **funtore rappresentato da A**, o funtore di Hom covariante in A:

$$\mathrm{Hom}_{\mathscr{C}}(A, -): \mathscr{C} \to \mathrm{Set}$$

che assegna a ogni oggetto B la classe $\mathrm{Hom}_{\mathscr{C}}(A, B)$ dei morfismi da A a B, e a ogni morfismo $f: B \to C$ la funzione di post-composizione:

$$\mathrm{Hom}_{\mathscr{C}}(A, f): \mathrm{Hom}_{\mathscr{C}}(A, B) \to \mathrm{Hom}_{\mathscr{C}}(A, C), \quad g \mapsto f \circ g$$

Analogamente si definisce il funtore controvariante $\mathrm{Hom}_{\mathscr{C}}(-, A): \mathscr{C}^{\mathrm{op}} \to \mathrm{Set}$ che, per pre-composizione, assegna a ogni oggetto B la classe $\mathrm{Hom}_{\mathscr{C}}(B, A)$.

Il lemma di Yoneda afferma allora:

Lemma (Yoneda, 1954). Per ogni categoria localmente piccola $\mathscr{C}$, ogni oggetto $A \in \mathrm{Ob}(\mathscr{C})$ e ogni funtore $F: \mathscr{C} \to \mathrm{Set}$, esiste una biiezione naturale:

$$\mathrm{Nat}(\mathrm{Hom}_{\mathscr{C}}(A, -), F) \cong F(A)$$

dove Nat(-, -) indica la classe delle trasformazioni naturali tra due funtori. Questa biiezione è naturale sia in A che in F.

La dimostrazione è sorprendentemente diretta. Data una trasformazione naturale α: $\mathrm{Hom}_{\mathscr{C}}(A, -) \Rightarrow F$, si valuta la sua componente sull'oggetto A:

$$\alpha_A : \mathrm{Hom}_{\mathscr{C}}(A, A) \to F(A)$$

e si applica al morfismo identità $\mathrm{id}_A \in \mathrm{Hom}_{\mathscr{C}}(A, A)$, ottenendo un elemento:

$$y(\alpha) := \alpha_A(\mathrm{id}_A) \in F(A)$$

Viceversa, dato un elemento $x \in F(A)$, si costruisce una trasformazione naturale β^x: $\mathrm{Hom}_{\mathscr{C}}(A, -) \Rightarrow F$ definendo la componente su un oggetto B come:

$$\beta^x_B : \mathrm{Hom}_{\mathscr{C}}(A, B) \to F(B), \quad f \mapsto F(f)(x)$$

La naturalità di β^x discende direttamente dalla funtorialità di F: per ogni $g: B \to C$ in $\mathscr{C}$,

$$F(g)(\beta^x_B(f)) = F(g)(F(f)(x)) = (F(g) \circ F(f))(x) = F(g \circ f)(x) = \beta^x_C(g \circ f)$$

che è esattamente la condizione di naturalità. La corrispondenza $y: \alpha \mapsto \alpha_A(\mathrm{id}_A)$ e la sua inversa $\beta: x \mapsto \beta^x$ costituiscono la biiezione cercata; che esse siano inverse l'una dell'altra è una verifica elementare:

$$y(\beta^x) = \beta^x_A(\mathrm{id}_A) = F(\mathrm{id}_A)(x) = \mathrm{id}_{F(A)}(x) = x$$
$$\beta^{y(\alpha)}_B(f) = F(f)(\alpha_A(\mathrm{id}_A)) = \alpha_B(\mathrm{Hom}_{\mathscr{C}}(A, f)(\mathrm{id}_A)) = \alpha_B(f \circ \mathrm{id}_A) = \alpha_B(f)$$

dove nell'ultimo passaggio si è usata la naturalità di α.

Il contenuto filosofico del lemma di Yoneda è straordinariamente denso, e vale la pena esplicitarlo con attenzione. Esso afferma che un oggetto A è completamente determinato — «conoscibile» in senso pieno — dalla totalità dei morfismi che altri oggetti intrattengono con esso, ossia dal funtore $\mathrm{Hom}_{\mathscr{C}}(A, -)$. Due oggetti A e A' sono isomorfi in $\mathscr{C}$ se e solo se i loro funtori di Hom sono naturalmente isomorfi come funtori in $\mathrm{Fun}(\mathscr{C}, \mathrm{Set})$. L'identità degli oggetti si legge nella struttura dei morfismi: chi un oggetto è coincide con come esso si rapporta a tutti gli altri oggetti della categoria.

Questo principio — che potremmo chiamare **olismo morfologico** — ha una risonanza filosofica che merita di essere articolata. Esso formalizza, in modo sorprendentemente preciso, la tesi strutturalista in filosofia della matematica: gli oggetti matematici non hanno una «natura intrinseca» che li individui indipendentemente dalle relazioni in cui entrano, ma sono esauriti dalle relazioni stesse. La posizione di Benacerraf — che i numeri naturali non sono oggetti

particolari, ma ruoli strutturali in qualsiasi progressione ω — riceve qui una formulazione tecnica rigorosa: il numero 2 *è* il funtore Hom(2, -) nella categoria dei numeri naturali ordinata.

Applicato alla filosofia della mente, il lemma di Yoneda suggerisce una tesi di grande interesse: il contenuto di uno stato mentale M è completamente determinato dall'insieme dei morfismi che gli altri stati mentali intrattengono con esso — dalle inferenze che portano a M, da quelle che da M si dipartono, dalle relazioni di compatibilità, esclusione, implicazione reciproca che M intrattiene con l'intera rete degli stati cognitivi del soggetto. Questa è, in forma rigorosa, la tesi che Brandom esprime nel suo inferenzialismo: il contenuto concettuale è il ruolo inferenziale. Non c'è nulla «oltre» la rete dei morfismi; l'oggetto *è* la sua posizione nel grafo morfologico.

VI. Dove il formalismo si arresta: la soglia del fenomenico

Ed è precisamente a questo punto che il progetto formale incontra il proprio limite interno — e il modo in cui lo incontra è, esso stesso, filosoficamente illuminante.

Il lemma di Yoneda afferma che ogni oggetto di una categoria è identificato dalla struttura delle sue relazioni morfologiche. Ma considera ora questa situazione: due stati mentali M_1 e M_2 che siano **morfologicamente indiscernibili** — cioè tali che $\mathrm{Hom}(M_1, -)$ e $\mathrm{Hom}(M_2, -)$ siano naturalmente isomorfi, che ogni altro stato stia in relazione con M_1 esattamente come sta con M_2 — risulterebbero, per il lemma di Yoneda, identici nella categoria $\mathfrak{M}$. Eppure sembra possibile immaginare che M_1 e M_2 differiscano qualitativamente, per come si presentano al soggetto: che abbiano lo stesso ruolo inferenziale e nondimeno una diversa qualità fenomenica.

Questo è precisamente il contenuto filosofico degli **argomenti degli invertiti qualitativi** di Ned Block e Sydney Shoemaker. Immagina che lo spettro di colori percepito da due individui sia sistematicamente invertito: ciò che io vedo quando guardo il rosso ha la qualità fenomenica che tu hai quando guardi il verde, e viceversa. Se tuttavia entrambi abbiamo imparato a usare i termini cromatici esattamente nello stesso modo — se tutte le nostre disposizioni inferenziali, comportamentali e linguistiche sono identiche —, le nostre categorie mentali cromatiche sono morfologicamente identiche nella categoria $\mathfrak{M}$. Il lemma di Yoneda le dichiarerebbe isomorfe. Eppure qualcosa, in questa affermazione, sembra mancare.

La questione si può precisare in termini tecnici. Il funtore di Yoneda:

$$y: \mathfrak{M} \to \mathrm{Fun}(\mathfrak{M}^{\{op\}}, \mathrm{Set}), \quad M \mapsto \mathrm{Hom}_{\mathfrak{M}}(-, M)$$

è **pienamente fedele**: iniettivo sui morfismi tra oggetti qualsiasi, nel senso che per ogni coppia $M, N \in \mathfrak{M}$,

$$y_{M,N}: \mathrm{Hom}_{\mathfrak{M}}(M, N) \to \mathrm{Nat}(\mathrm{Hom}_{\mathfrak{M}}(-, M), \mathrm{Hom}_{\mathfrak{M}}(-, N))$$

è una biiezione. Ciò significa che la struttura categoriale di $\mathfrak{M}$ si incorpora perfettamente nella categoria dei prefasci: nessuna relazione morfologica va perduta nella transizione. Un funtore pienamente fedele è un'**immersione** categoriale; esso garantisce che la struttura morfologica di $\mathfrak{M}$ sia completamente preservata nell'immagine.

Ma la fedeltà del funtore di Yoneda non implica che *tutto* ciò che è reale negli oggetti di $\mathfrak{M}$ sia di natura morfologica. Essa implica soltanto che, *se* due oggetti differiscono, questa differenza si manifesta morfologicamente. L'argomento degli invertiti qualitativi afferma precisamente che la differenza qualitativa tra M_1 e M_2 potrebbe essere una differenza *non morfologica* — che sfugge alla rete dei morfismi — e che dunque non verrebbe rilevata dal funtore di Yoneda.

Formalmente: sia $\mathfrak{M}_\Phi$ la categoria degli stati mentali di un soggetto Φ e $\mathfrak{M}_\Psi$ quella di un soggetto Ψ con spettro cromatico invertito. Il fatto che le due categorie siano equivalenti — che esista un'equivalenza di categorie $\mathfrak{M}_\Phi \simeq \mathfrak{M}_\Psi$ — non esclude che esista una differenza tra i due sistemi che non sia catturata dalla struttura categoriale. Tale differenza sarebbe, per definizione, di tipo **extra-morfologico**.

Il problema difficile della coscienza trova qui una formulazione categoriale esatta: esso è la congettura che la struttura di $\mathfrak{M}$ non sia completa rispetto alle proprietà rilevanti degli stati mentali — che esista un «residuo» che nessun funtore da $\mathfrak{M}$ verso una categoria di strutture puramente formali riesce a preservare. Questo residuo è ciò che Chalmers chiama *qualia*, ciò che Nagel indica come il «what it is like», ciò che la tradizione fenomenologica denomina *Erlebnis* nel senso di Husserl o *Leib* nel senso di Merleau-Ponty.

Notare la precisione del punto di arresto: il formalismo categoriale non collassa di fronte al problema fenomenico, non lo dissolve come pseudo-problema né lo aggira con una ridefinizione operazionale. Lo *situa* con esattezza: il problema si trova esattamente al confine tra la struttura dei morfismi — che è tutto ciò su cui la teoria delle categorie può parlare — e la presunta struttura non-morfologica dell'esperienza. L'argomento degli invertiti qualitativi, nella lettura categoriale, è l'affermazione che $\mathfrak{M}$ ammette automorfismi non triviali — isomorfismi con se stessa che non sono l'identità — che preservano la struttura categoriale ma non le qualità fenomeniche. L'isomorfismo categoriale non è sufficiente per l'identità fenomenica.

V.1 — Riformulare il gap esplicativo: da enigma filosofico a problema inverso

Il *gap esplicativo* di Joseph Levine e la *difficult question* di David Chalmers occupano, nel panorama della filosofia della mente contemporanea, il posto che il problema del moto retrograde dei pianeti occupava nell'astronomia tolemaica: sono l'anomalia che ogni teoria deve confrontare, pena l'irrilevanza. La domanda è brutalmente semplice:

perché certi processi fisici — certi pattern di attivazione neurale, certe strutture computazionali — sono accompagnati da esperienza soggettiva, da quel *sapere che cos'è* che Nagel ha eternizzato nell'immagine del pipistrello? E perché la descrizione fisica, per quanto completa e dettagliata, sembra sempre lasciare aperta la possibilità logica che quella stessa struttura fisica esista senza alcuna esperienza — il che è l'affermazione di concepibilità degli zombie filosofici di Chalmers?

La teoria geometrica proposta in questa monografia non pretende di eliminare questa difficoltà: sarebbe disonesto sostenerlo. Ciò che si propone è qualcosa di più modesto ma anche più produttivo: una *riformulazione* del gap esplicativo in termini matematicamente precisi, tale da trasformarlo da un enigma verbale a un problema tecnico ben posto — o, se si preferisce, da mostrare con esattezza dove la precisione tecnica si ferma e dove il gap rimane genuino.

La riformulazione è la seguente. Si è sostenuto nelle parti precedenti che il contenuto intenzionale è determinato dalla struttura metrica intrinseca della varietà rappresentazionale. Ora, ciò che gli scienziati cognitivi e i neuroscienziati misurano sono le *attivazioni neurali*: vettori in $\mathbb{R}^N$ che sono punti sulla varietà immersa nello spazio di attivazione. Questi vettori sono la descrizione *estrinseca* della varietà — la sua rappresentazione nell'ambiente fisico misurabile. La struttura *intrinseca* — la metrica di Fisher, la curvatura, la topologia — deve essere *ricostruita* a partire da questi dati estrinseci.

Questo è precisamente il *problema inverso dell'embedding*: dato un insieme di punti nello spazio di attivazione (la varietà immersa), ricostruire la metrica intrinseca della varietà sottostante. Il problema diretto — dato una varietà astratta con una certa metrica, trovare un embedding nello spazio euclideo — ha una risposta classica: il *teorema di embedding isometrico di Nash* afferma che qualsiasi varietà riemanniana compatta di dimensione n può essere isometricamente immersa in $\mathbb{R}^N$ per N sufficientemente grande (in particolare, per $N = n(3n+11)/2$ nel caso liscio). Ma il problema inverso — ricostruire la metrica intrinseca da un embedding osservato — è in generale mal posto e richiede ipotesi aggiuntive di regolarità.

Il gap esplicativo, riformulato in questi termini, è il seguente: anche se si disponesse di una misura perfetta di tutte le attivazioni neurali (il dato estrinseco), la ricostruzione della metrica intrinseca — che costituisce il contenuto secondo la teoria proposta — è un problema mal posto che ammette molteplici soluzioni compatibili con i dati. Detto altrimenti: la stessa struttura di attivazioni neurali può essere compatibile con strutture di contenuto intrinsecamente diverse, e i dati fisici da soli non possono discriminare tra queste possibilità. Questo è il gap geometrico: la descrizione fisica è necessariamente una proiezione di dimensionalità ridotta di una struttura intrinseca che la trascende.

V.2 — Il teorema di Whitney e la sua portata filosofica

Per rendere questo argomento preciso, occorre richiamare i risultati classici sulla relazione tra varietà astratte e loro embedding. Il *teorema di embedding di Whitney* (1944) nella sua forma forte afferma:

Qualsiasi varietà differenziabile compatta di dimensione n può essere immersa (senza auto-intersezioni) in $\mathbb{R}^{2n}$.

Questo risultato garantisce l'esistenza di un embedding, ma non dice nulla sull'unicità né sulla possibilità di ricostruire la varietà dall'embedding. Il problema di ricostruzione — l'inverso del teorema di Whitney — è il seguente: dato un insieme $S \subset \mathbb{R}^{2n}$ che è l'immagine di un embedding di una varietà n-dimensionale, ricostruire la varietà e la sua metrica intrinseca.

In generale, questo problema non ha soluzione unica. Due varietà riemanniane non isometriche possono avere embedding in $\mathbb{R}^{2n}$ le cui immagini sono arbitrariamente vicine nella topologia di Hausdorff. Il *teorema di rigidità di Cohn-Vossen* fornisce condizioni sotto le quali un embedding è unico: per le superfici compatte convesse (curvatura di Gauss positiva), l'embedding in $\mathbb{R}^3$ è determinato a meno di isometrie rigide del piano. Ma per varietà di dimensione superiore e curvatura arbitraria — che è il caso rilevante per le varietà neurali — non esiste un risultato generale di rigidità.

La misura dell'indeterminazione è catturata dal concetto di *flessibilità isometrica*: l'insieme delle deformazioni della metrica intrinseca compatibili con un dato embedding. Per una varietà n-dimensionale immersa in $\mathbb{R}^{2n}$, lo spazio delle deformazioni isometriche compatibili ha in generale dimensione infinita — è un insieme di infiniti oggetti geometricamente distinti che producono lo stesso dato estrinseco. Nella tradizione della geometria differenziale, questo è il *teorema di flessibilità di Nash-Kuiper*: qualsiasi embedding liscio può essere C^1-approssimato arbitrariamente bene da embedding isometrici di qualsiasi varietà con la stessa topologia.

La portata filosofica di questi risultati è precisa. Essi dimostrano che la transizione dalla descrizione estrinseca (le attivazioni neurali misurabili) alla struttura intrinseca (la metrica del contenuto) è un'inferenza *sottodeterminata*: i dati fisici non determinano univocamente il contenuto. Questo non è un gap empirico contingente — non è che gli strumenti di misura attuali siano troppo grossolani — ma è un gap strutturale che sussiste anche nell'ipotetica situazione di misure perfette. La fisica non determina la metrica del contenuto nello stesso senso in cui, per fare un'analogia con il linguaggio, la serie di suoni fisici prodotta da una frase non determina univocamente il suo contenuto proposizionale senza un contesto interpretativo.

V.3 — Una proposta risolutiva parziale: il vincolo del principio di massima curvatura

Nonostante l'indeterminazione di principio appena discussa, vi sono vincoli aggiuntivi che restringono l'insieme delle metriche intrinseche compatibili con un dato embedding neurale. Il più interessante, che qui si propone in modo originale, è il *principio di massima curvatura informazionale*, ispirato al principio di massima entropia di Jaynes ma distinto da esso nella struttura.

Il principio si enuncia come segue. Tra tutte le metriche intrinseche compatibili con un dato embedding estrinseco (le attivazioni neurali), la metrica realizzata è quella che *massimizza la curvatura scalare* della varietà rappresentazionale, soggetta al vincolo che la varietà rimanga completa geodetica (senza buchi o singolarità). Formalmente:

$$g^* = \text{argmax}_{g \in \mathcal{G}(\varepsilon)} \int_{\mathcal{V}} R(g) \cdot dVol_g$$

dove $\mathcal{G}(\varepsilon)$ è l'insieme di tutte le metriche riemanniane la cui embedding isometrico è a distanza di Hausdorff minore di ε dall'embedding osservato, $R(g)$ è la curvatura scalare di Ricci della varietà $(\mathcal{V}, g)$, e $dVol_g$ è la forma di volume indotta da g.

La curvatura scalare R è la traccia del *tensore di Ricci* Ric_{ij}, che a sua volta è la traccia parziale del tensore di curvatura di Riemann:

$$Ric_{ij}(\theta) = R^k{}_{ikj}(\theta) = \Sigma \; R^k{}_{ikj}(\theta)$$

$$R(\theta) = g^{ij} \, Ric_{ij}(\theta)$$

La curvatura scalare misura il grado complessivo di non-planarità della varietà in un punto: in uno spazio piatto $R = 0$ ovunque; in uno spazio sferico di raggio r, $R = n(n-1)/r^2$; in uno spazio iperbolico di curvatura $-\kappa$, $R = -n(n-1)\kappa$.

La giustificazione del principio di massima curvatura si articola su due livelli. Sul piano *informatico-computazionale*, una varietà a curvatura elevata ha una capacità di discriminazione locale più alta: le sfere geodetiche crescono meno rapidamente (spazi a curvatura positiva) o le geodetiche divergono più rapidamente (spazi a curvatura negativa), in entrambi i casi rendendo gli stati vicini più facilmente distinguibili. Massimizzare la curvatura significa dunque massimizzare la *resolving power* rappresentazionale del sistema cognitivo — la sua capacità di fare distinzioni fini nello spazio dei contenuti. Sul piano *evolutivo-funzionale*, un sistema cognitivo che massimizza la curvatura del suo spazio rappresentazionale è un sistema che estrae il massimo di informazione discriminatoria dall'ambiente con il minimo di risorse rappresentazionali.

Sul piano *fenomenologico*, la curvatura massima corrisponde alla *ricchezza qualitativa* dell'esperienza. Uno spazio piatto — curvatura nulla — sarebbe uno spazio in cui tutti i percorsi tra stati sono equivalenti, in cui non vi è struttura gerarchica né relazioni di dipendenza tra contenuti: un'esperienza fenomenologicamente povera o nulla. Uno spazio altamente curvo è uno spazio in cui le geodetiche si intersecano, si avvolgono, si separano in modi complessi — un'esperienza ricca di sfumature, strutture e

relazioni. Sebbene questo argomento fenomenologico non sia dimostrativo, esso suggerisce una connessione tra curvatura rappresentazionale e ricchezza qualitativa dell'esperienza che merita indagine empirica.

V.4 — La olonomia dei processi cognitivi e il legame temporale della coscienza

Si è introdotta nella Parte Seconda la nozione di olonomia di una connessione: il fatto che il trasporto parallelo di un vettore lungo un loop chiuso produce in generale una rotazione del vettore, e che il gruppo di olonomia misura l'insieme di tutte le possibili rotazioni così generate. Occorre ora sviluppare il significato di questa struttura per la teoria del contenuto temporale della coscienza.

Il problema della *temporalità della coscienza* è uno dei più profondi nella fenomenologia e nella filosofia della mente: come fa la coscienza a tenere insieme il presente vivente con la ritenzione del passato immediato e la protenzione del futuro immediato (per usare il linguaggio husserliano della *Zeitbewusstsein*)? Come si costruisce il senso di continuità e di flusso dell'esperienza a partire da stati neurali discreti o quasi-discreti? Quale struttura fisica o computazionale è responsabile del fatto che l'esperienza cosciente ha una *forma temporale* — è un flusso, non una successione di istanti isolati?

La risposta che si propone introduce la olonomia come struttura matematica fondamentale per il legame temporale. Si consideri un processo cognitivo come la comprensione di una frase — un processo che si dispiega nel tempo e che alla sua conclusione sembra essere «tenuto insieme» in un'unità di senso. Nella varietà rappresentazionale, questo processo corrisponde a una curva $\gamma\colon [0, T] \to \mathcal{V}$ che parte dallo stato iniziale $\gamma(0)$ e arriva allo stato finale $\gamma(T)$. Se $\gamma(0) = \gamma(T)$ — se il processo è ciclico, per esempio se si tratta di un ritmo melodico che si ripete — allora γ è un loop chiuso, e il trasporto parallelo lungo γ è l'elemento di olonomia $H_\gamma \in \mathrm{Hol}(\gamma(0), \nabla)$.

L'elemento di olonomia H_γ non è, in generale, la rotazione identità. Esso trasforma il vettore tangente iniziale v_0 in un vettore $v_T = H_\gamma(v_0)$ che è ruotato rispetto a v_0. Questa rotazione è la *memoria geometrica* del processo: essa codifica, in forma compatta, l'effetto cumulativo di tutte le variazioni di curvatura incontrate lungo il percorso. La olonomia è dunque la struttura matematica che permette a un processo cognitivo di «tornare allo stesso stato» e tuttavia essere *arricchito* dall'esperienza del percorso: lo stato cognitivo finale $\gamma(T) = \gamma(0)$ è geometricamente identico allo stato iniziale, ma la struttura tangente — i vettori nello spazio tangente — è stata trasformata dall'olonomia del percorso.

Questo suggerisce la seguente interpretazione: il *legame temporale* della coscienza — il fatto che gli stati cognitivi distanti nel tempo sono tenuti insieme in un'unità di esperienza — corrisponde alla struttura del gruppo di olonomia $\mathrm{Hol}(\mathcal{V}, \nabla)$. Un gruppo di olonomia triviale (solo la rotazione identità) corrisponde a un'esperienza priva di legame temporale, in cui ogni stato è completamente isolato dagli altri. Un gruppo di olonomia ricco — un gruppo di Lie non abeliano, per esempio — corrisponde a un'esperienza in cui i processi temporali lasciano tracce strutturali

profonde nello spazio tangente, ovvero a una coscienza con memoria implicita ricca e legame temporale robusto.

Il teorema di Berger classifica i possibili gruppi di olonomia per le varietà riemanniane semplicemente connesse irriducibili: essi appartengono a una lista finita che comprende SO(n) (il caso generico), U(n/2) (le varietà kähleriane), SU(n/2) (le varietà Calabi-Yau), Sp(n/4) (le varietà iperquaternioniche), e i casi eccezionali G_2 e Spin(7). Questa classificazione è notevole: dice che i tipi di legame temporale possibili, nella teoria geometrica, non sono arbitrari ma appartengono a una lista discreta di strutture algebriche. Diverse strutture di olonomia corrisponderebbero a diverse *modalità di coscienza temporale* — e ciò suggerisce la possibilità di una tassonomia dei modi di esperienza temporale basata sulla classificazione di Berger, il che sarebbe un risultato genuinamente inaspettato dell'approccio geometrico.

La connessione con la *sincronizzazione neurale oscillatoria* — uno dei meccanismi candidati per il legame temporale nel cervello — può essere esplicitata come segue. Le oscillazioni gamma (30-100 Hz) che si osservano durante l'elaborazione percettiva e che sono state proposte da Crick, Koch e poi da Singer come possibile substrato del legame temporale corrispondono, nella teoria geometrica, a loop quasi-periodici sulla varietà rappresentazionale. La sincronizzazione di oscillazioni in aree diverse corrisponde all'allineamento delle fasi di loop in diversi sottoinsieme della varietà — un allineamento che produce un gruppo di olonomia comune, e quindi un legame temporale condiviso tra le aree. La desincronizzazione — l'assenza di gamma-binding — corrisponde a gruppi di olonomia incoerenti tra aree, e dunque a una mancanza di legame temporale, compatibile con la disgregazione dell'esperienza cosciente nell'anestesia generale o in certi stati dissociativi.

V.5 — Il fantasma della circolarità e la risposta

Una critica naturale all'intero impianto teorico di questa monografia è quella della *circolarità esplicativa*. Si sostiene che il contenuto mentale è determinato dalla struttura geometrica della varietà rappresentazionale. Ma come si individua la varietà rappresentazionale se non attraverso il contenuto degli stati cognitivi? La struttura geometrica presuppone di sapere quali stati cognitivi sono «simili» e quali «diversi», il che equivale già a presupporre una struttura di contenuto. Si starebbe dunque definendo il contenuto in termini di una struttura che a sua volta presuppone il contenuto.

Questa obiezione è seria e merita una risposta articolata. La risposta si articola in tre mosse distinte.

Prima mossa: la varietà rappresentazionale non è identificata attraverso il contenuto ma attraverso la *dinamica*. Le traiettorie degli stati cognitivi nello spazio di attivazione — come certi stati tendono a seguire certi altri, come certi stati si attraggono reciprocamente, come certi stati sono stabili e certi instabili — determinano la struttura geometrica della varietà indipendentemente da qualsiasi interpretazione semantica. La metrica di Fisher emerge dalle distribuzioni di transizione, non dai

contenuti. È dunque possibile, in linea di principio, determinare la struttura geometrica della varietà rappresentazionale da misure puramente dinamiche (registrazioni elettrofisiologiche, imaging funzionale) senza presupporre alcun contenuto.

Seconda mossa: anche se vi fosse un certo grado di interdipendenza tra la determinazione della struttura geometrica e la attribuzione di contenuto, ciò non implica circolarità viziosa ma *coerenza sistemica*. Il contenuto e la struttura geometrica si determinano mutuamente secondo un sistema di equazioni di fissazione del punto, e la domanda rilevante è se questo sistema abbia soluzioni e se esse siano stabili. Nella teoria dei punti fissi — il teorema di Brouwer e le sue generalizzazioni — sistemi di equazioni apparentemente circolari hanno spesso soluzioni uniche e stabili. La circolarità apparente è, in questo caso, la struttura di un sistema autoreferenziale ben posto, non un difetto logico.

Terza mossa: si può ribaltare la critica e mostrare che essa si applica parimenti a tutte le teorie rivali. Qualsiasi teoria che riduce il contenuto a relazioni causali deve presupporre quale tra le molteplici cause di uno stato sia quella «rilevante» — e questa scelta non può essere fatta senza fare riferimento al contenuto che si intende spiegare. Qualsiasi teoria funzionalista deve presupporre quali transizioni tra stati siano quelle «funzionalmente rilevanti» — e anche questa scelta implica una precomprensione del contenuto. Il problema della circolarità non è specifico della teoria geometrica: esso è una condizione trascendentale di qualsiasi progetto di naturalizzazione del contenuto, e la teoria geometrica lo affronta con gli strumenti più precisi disponibili.

VII. La metrica di Fisher-Rao e la geometria dello spazio dei contenuti

Il passaggio dalla teoria delle categorie alla geometria dell'informazione apre una prospettiva complementare che approfondisce la diagnostica del limite formale. Si consideri lo spazio dei modelli probabilistici su un insieme di dati sensoriali X. Una famiglia parametrica di distribuzioni:

$$\mathscr{S} = \{\, p(x \mid \theta) : \theta \in \Theta \subseteq \mathbb{R}^n \,\}$$

dove Θ è un aperto di $\mathbb{R}^n$, definisce una **varietà statistica**: uno spazio differenziale n-dimensionale in cui ogni punto θ corrisponde a un modello probabilistico. Su questa varietà si può definire una metrica riemanniana naturale — la **metrica di Fisher-Rao** — attraverso il suo tensore metrico:

$$g_{ij}(\theta) = E_{p(x|\theta)}\left[(\partial/\partial\theta_i) \log p(x|\theta) \cdot (\partial/\partial\theta) \log p(x|\theta) \right]$$
$$= \int p(x|\theta) \cdot (\partial/\partial\theta_i \log p(x|\theta)) \cdot (\partial/\partial\theta \log p(x|\theta))\, dx$$

ovvero, equivalentemente, usando il fatto che $\int p(x|\theta)\, dx = 1$ per derivare la forma equivalente:

$$g_{ij}(\theta) = -E_{p(x|\theta)}\left[\partial^2/\partial\theta_i\partial\theta \log p(x|\theta) \right]$$

La matrice $g_{ij}(\theta)$ è la **matrice di informazione di Fisher**, e il fatto fondamentale — dimostrato da Chentsov nel 1972 — è che essa è l'unica metrica riemanniana su $\mathscr{S}$ che sia *invariante per riparametrizzazioni sufficienti* della statistica, ossia per trasformazioni che non perdono informazione. In questo senso la metrica di Fisher-Rao non è una scelta arbitraria: è la struttura geometrica intrinseca dello spazio dei modelli statistici.

Su questa varietà riemanniana si può definire la connessione di Levi-Civita associata a g_{ij}, e quindi il tensore di curvatura di Riemann:

$$R_i{}^{l} = \partial \Gamma^l_i - \partial_i \Gamma^l + \Gamma^l \Gamma^m_i - \Gamma^l_i \Gamma^m$$

dove i simboli di Christoffel sono:

$$\Gamma^l_i = (1/2)\, g^{lm} (\partial_i g_{jm} + \partial g_{im} - \partial g_{ij})$$

La curvatura di Riemann misura il grado in cui lo spazio dei modelli si discosta dalla piattezza euclidea: una varietà statistica piatta (curvatura nulla) corrisponde a una famiglia esponenziale, nella quale la distanza tra modelli è «regolare» nel senso che il trasporto parallelo è globalmente definito e indipendente dal percorso.

La connessione tra geometria dell'informazione e intenzionalità si articola come segue. Se si identifica lo spazio dei parametri θ con lo spazio degli stati intenzionali — ogni θ corrisponde a un «modo di vedere il mondo», a un contenuto intenzionale — la metrica di Fisher-Rao misura la distinguibilità tra stati intenzionali vicini: $g_{ij}(\theta)$ esprime quanto rapidamente la distribuzione sulle esperienze sensoriali cambia al variare del contenuto intenzionale. Due stati fenomenicamente diversi ma inferenzialmente indistinguibili corrisponderanno, in questa geometria, a punti vicini rispetto alla metrica di Fisher-Rao ma distanti rispetto a una metrica fenomenica ipotetica — se tale metrica esiste.

Il problema difficile, riformulato geometricamente, diventa allora: **esiste una varietà fenomenica** $\mathscr{P}$ dotata di una propria struttura metrica — misurante la distanza «qualitativa» tra esperienze — tale che la mappa naturale $\mathscr{P} \to \mathscr{S}$ (che associa a ogni qualità fenomenica il modello statistico corrispondente) sia una mappa non isometrica? Se la risposta è affermativa, allora la struttura geometrica di $\mathscr{P}$ non è riducibile a quella di $\mathscr{S}$, e il residuo fenomenico persiste come un invariante geometrico che la teoria probabilistica non cattura. Se è negativa — se la mappa è un'isometria — allora il contenuto fenomenico è completamente determinato dalla struttura statistica dell'inferenza, e il problema difficile dissolve (o si rivela mal posto).

Non si è in grado, al momento, di decidere empiricamente tra queste due possibilità. Ma la precisione con cui il problema si può porre in termini geometrici è già un risultato: essa mostra che il problema difficile non è semplicemente un residuo di confusione concettuale, ma una questione tecnica la cui risposta richiede la

costruzione di una geometria dell'esperienza fenomenica — impresa che la neuroscienza attuale, nonostante i tentativi di Tononi con la sua teoria dell'informazione integrata, non ha ancora portato a compimento in modo soddisfacente.

Il funtore dimenticante, l'informazione integrata e i limiti del riduzionismo strutturale

---

III. Varietà Neurali e la Struttura Geometrica delle Rappresentazioni

---

III.1 — L'ipotesi della varietà neurale: dagli esperimenti alla teoria

Uno degli sviluppi più fecondi delle neuroscienze computazionali degli ultimi tre decenni è la progressiva consapevolezza che l'attività dei circuiti neurali non è distribuita uniformemente e caoticamente nello spazio delle configurazioni possibili, ma è vincolata a strutture di dimensione bassa immerse in uno spazio di attivazione ad alta dimensionalità. Questa intuizione — formulata inizialmente in modo informale da ricercatori come Seung e Lee nel contesto delle mappe sensoriali, e poi precisata tecnicamente dalla teoria delle *neural manifolds* — è diventata uno dei paradigmi guida della neuroscienza teorica contemporanea.

L'osservazione empirica di partenza è la seguente. Un'area corticale come V1, il cortex visivo primario, contiene dell'ordine di 10^8 neuroni, ciascuno dei quali può in principio assumere un continuo di livelli di attivazione. Lo spazio delle configurazioni teoricamente possibili ha dunque dimensione dell'ordine di 10^8 — un oggetto di dimensionalità inimmaginabilmente alta. Eppure, quando si registra l'attività di un sottoinsieme rappresentativo di neuroni di V1 mentre un animale guarda stimoli visivi naturali, si scopre che i vettori di attivazione osservati non riempiono lo spazio ad alta dimensionalità in modo uniforme: essi giacciono, con buona approssimazione, su una sottovarietà di dimensione molto bassa — dell'ordine di 10^2 al più. Gli studi di Cadieu e colleghi su V1 del macaco, e quelli di DiCarlo e Cox su IT (cortex infertemporale), hanno mostrato che la dimensione intrinseca delle rappresentazioni sensoriali è compresa tipicamente tra alcune decine e alcune centinaia, ben al di sotto della dimensione nominale dello spazio di attivazione.

Questo fenomeno — la *compressione rappresentazionale* — non è triviale. Esso non si limita a dire che le aree corticali non usano tutta la loro capacità teorica in un dato momento: dice qualcosa di molto più forte, ovvero che la topologia e la geometria dello spazio effettivo delle attivazioni è fondamentalmente diversa da quella dello spazio nominale. La varietà neurale ha una sua curvatura, una sua topologia globale,

una sua struttura di connessione — e queste proprietà geometriche intrinseche sembrano essere correlate in modo sistematico con le proprietà degli stimoli rappresentati.

Il risultato più emblematico in questa direzione è quello di Donoghue, Shenoy e colleghi sulle popolazioni di neuroni motori: le traiettorie degli stati di attivazione durante movimenti del braccio giacciono su una varietà toroidale di bassa dimensione, e la struttura topologica di questo toro — i suoi cicli non contraibili — corrisponde esattamente alla struttura dei gradi di libertà del movimento. La topologia della varietà neurale *codifica* la topologia dello spazio dei compiti motori. Analogamente, Gardner e colleghi hanno recentemente dimostrato che le cellule di luogo nell'ippocampo del topo si organizzano in una struttura toroidale che riflette la topologia dell'ambiente fisico navigato dall'animale. La geometria del mondo si proietta, attraverso l'esperienza, sulla geometria dello spazio rappresentazionale.

III.2 — Stima della dimensione intrinseca e le sue ambiguità

Prima di procedere, è necessario essere precisi riguardo a cosa si intende per «dimensione intrinseca» di una varietà neurale, perché questo concetto è soggetto a molteplici interpretazioni tecniche con implicazioni filosofiche diverse.

La nozione più elementare è quella di *dimensione di correlazione*, introdotta da Grassberger e Procaccia nel contesto dei sistemi dinamici. Data una nuvola di punti $X = \{x_1, ..., x_N\}$ in $\mathbb{R}^D$, la dimensione di correlazione è definita come:

$$d_{corr} = \lim_{r \to 0} \log C(r) / \log r$$

dove $C(r)$ è la *integrale di correlazione*:

$$C(r) = \lim_{N \to \infty} (2 / N(N-1)) \cdot \#\{(i,j) : \|x_i - x\| < r\}$$

Se i dati giacciono su una varietà liscia di dimensione d, allora per r sufficientemente piccolo $C(r) \propto r^d$, e la dimensione di correlazione converge a d. Questa stima è robusta a trasformazioni lisce dei dati, ma è numericamente instabile in alta dimensione e richiede campioni molto grandi.

Una stima più stabile, ampiamente usata nelle neuroscienze, è quella basata sull'*analisi delle componenti principali* (PCA): si calcola la matrice di covarianza dei vettori di attivazione, se ne calcolano gli autovalori $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$, e si definisce la dimensione intrinseca come il numero di autovalori necessari per spiegare una frazione prefissata (tipicamente 90-95%) della varianza totale. Questa stima corrisponde alla dimensione del sottospazio euclideo lineare di migliore approssimazione, e ha il difetto di non cogliere la curvatura non lineare della varietà.

La stima non lineare più raffinata, proposta da Levina e Bickel e poi migliorata da vari autori, si basa sull'idea che, in una varietà di dimensione d, il numero di vicini contenuti in una palla di raggio r scala come r^d. Si stima localmente la dimensione usando le distanze tra vicini più prossimi:

$$\hat{d}_i = \left[\frac{1}{k} \sum_{j=1}^{k} \log\left(T_{ik} / T_{ij}\right) \right]^{-1}$$

dove T_{ij} è la distanza dal punto x_i al suo j-esimo vicino più prossimo. La dimensione globale è poi stimata come media delle stime locali $\hat{d}_i$. Questa procedura — chiamata stima *MLE della dimensione intrinseca* — è consistente sotto ipotesi di regolarità sulla varietà e converge alla dimensione topologica vera quando $N \to \infty$.

Perché questa discussione tecnica è filosoficamente rilevante? Perché le diverse nozioni di dimensione intrinseca catturano aspetti diversi della struttura geometrica della varietà rappresentazionale, e quindi aspetti diversi del contenuto mentale secondo la teoria proposta. La dimensione di correlazione cattura la struttura *scalare* — come la varietà riempie lo spazio a diverse scale. La dimensione PCA cattura la struttura *lineare* — i gradi di libertà della variazione lineare. La stima MLE cattura la struttura *locale* non lineare. Una teoria del contenuto basata sulla geometria della varietà deve specificare quale di queste strutture — o quale combinazione di esse — è rilevante per il contenuto, e questa scelta non è arbitraria ma dipende da quale aspetto del contenuto si vuole catturare.

La proposta di questo lavoro è che il contenuto intenzionale in senso pieno richiede la struttura *globale* della varietà — la sua topologia, la sua metrica, la sua curvatura — e non può essere ridotto a nessuna delle sue proiezioni locali o lineari. Questo implica che la PCA, pur essendo un utile strumento esplorativo, è in linea di principio inadeguata per catturare il contenuto: due spazi di attivazione con la stessa struttura lineare ma topologie globali diverse — uno sferico e uno toroidale, per esempio — avranno contenuti diversi nel senso rilevante per l'intenzionalità.

III.3 — Metrica intrinseca degli spazi di attivazione e la connessione con Fisher

Il passo cruciale consiste ora nel congiungere la geometria differenziale degli spazi statistici (Parte Seconda) con la geometria empirica delle varietà neurali. Il collegamento è fornito dalla seguente osservazione, che si articola in due direzioni.

Prima direzione: ogni stato di attivazione neurale può essere interpretato come una distribuzione di probabilità. In una descrizione probabilistica dell'attività neurale — standard nei modelli computazionali dalla teoria della codifica efficiente di Barlow in poi — ogni neurone i ha un tasso di scarica medio $r_i(\theta)$ che dipende dagli stimoli θ, e l'attività dell'intera popolazione può essere descritta come una distribuzione di Poisson (o, più accuratamente, come una distribuzione di variabili casuali dipendenti con struttura di correlazione non triviale). La distribuzione della risposta dell'intera popolazione a uno stimolo θ è:

$$p(r_1, r_2, ..., r\_N ; \theta) = \prod_i p\_i(r_i ; \theta) \cdot (1 + \Sigma\_\{i<j\} \Delta_i\square(\theta) \cdot (r_i - \bar{r_i})(r - \bar{r}\square) + ...)$$

dove il secondo termine tiene conto delle correlazioni tra neuroni e $\bar{r_i} = r_i(\theta)$ è il tasso medio. La metrica di Fisher di questa distribuzione rispetto ai parametri θ — che descrivono gli stimoli nel mondo — è:

$$g\hat{}(pop)\_\{kl\}(\theta) = \Sigma_i \, (1/r_i(\theta)) \cdot (\partial r_i/\partial\theta^k) \cdot (\partial r_i/\partial\theta^l) - \Sigma\_\{i,j\} \, C_i\square(\theta) \cdot (\partial \log r_i/\partial\theta^k) \cdot (\partial \log r /\partial\theta^l)$$

dove $C_i\square(\theta)$ è il tensore di correlazione tra i neuroni i e j. Il primo termine è il contributo dei singoli neuroni — si somma il quadrato del gradiente di ogni curva di risposta, pesato con l'inverso della varianza di Poisson — mentre il secondo termine è la *correzione per le correlazioni*.

Questo risultato — una versione raffinata del classico *Fisher information of a neural population* — ha una lettura geometrica diretta: la metrica di Fisher sullo spazio degli stimoli θ è indotta dalla mappa di codifica $r(\theta)$: $\Theta \to \mathbb{R}\hat{}N$. Più precisamente, se si considera la mappa $f: \Theta \to \mathbb{R}\hat{}N$ definita da $f(\theta) = (r_1(\theta), ..., r\_N(\theta))$, il *pullback* della metrica euclidea standard di $\mathbb{R}\hat{}N$ attraverso f è:

$$(f^*g\_euclid)\_\{kl\}(\theta) = \Sigma_i \, (\partial r_i/\partial\theta^k) \cdot (\partial r_i/\partial\theta^l)$$

che è il contributo dei singoli neuroni alla metrica di Fisher (trascurando le correlazioni). In altri termini, la metrica di Fisher sullo spazio degli stimoli è la metrica indotta dalla mappa di codifica neurale: è la metrica che misura le distanze nello spazio degli stimoli come vengono «percepite» dalla popolazione neurale. Due stimoli θ e θ' sono distanti nella metrica di Fisher nella misura in cui producono pattern di attivazione facilmente distinguibili.

Seconda direzione: si può invertire la prospettiva. Invece di vedere la metrica di Fisher come una metrica sullo spazio degli *stimoli* indotta dalla mappa di codifica, la si può vedere come una metrica sullo spazio delle *attivazioni* stesse. Dato un vettore di attivazione r, si può definire la varietà rappresentazionale come lo spazio parametrico $\{r(\theta) : \theta \in \Theta\} \subset \mathbb{R}\hat{}N$ — la varietà immagine della mappa di codifica. Questa varietà ha una metrica intrinseca — la metrica di Fisher pullback — che è indipendente dalla parametrizzazione θ scelta e dipende solo dalla struttura della mappa di codifica.

La tesi centrale della teoria proposta può ora essere enunciata in termini tecnici precisi: il *contenuto intenzionale* di uno stato di attivazione r è la *classe di isometria* della varietà rappresentazionale in un intorno geodetico sufficientemente ampio di r rispetto alla metrica di Fisher intrinseca. Due stati cognitivi — in qualsiasi substrato realizzati — hanno lo stesso contenuto intenzionale se e solo se esistono intorni geodetici isometrici nel senso della metrica di Fisher intrinseca.

III.4 — Embedding iperbolico e struttura gerarchica del contenuto

Un risultato della geometria differenziale che ha ricevuto attenzione crescente nel contesto dell'apprendimento automatico e che qui assume un significato teorico autonomo è la capacità degli spazi iperbolici di *embeddarsi* naturalmente strutture gerarchiche con distorsione logaritmica.

Lo spazio iperbolico di dimensione n con curvatura $-\varkappa$, denotato $H^n_{-\varkappa}$, è una varietà riemanniana completa, semplicemente connessa e a curvatura settoriale costante uguale a $-\varkappa$. Nel modello del disco di Poincaré, i punti sono i punti del disco unitario aperto $D^n = \{x \in \mathbb{R}^n : \|x\| < 1\}$, e la metrica è:

$$ds^2 = 4\varkappa / (1 - \|x\|^2)^2 \cdot \Sigma_i (dx^i)^2$$

La proprietà fondamentale degli spazi iperbolici che li rende adatti all'embedding di gerarchie è che il volume di una palla di raggio r cresce come $e^{(n-1)\sqrt{\varkappa} \cdot r}$, dunque esponenzialmente nel raggio, anziché polinomialmente come nello spazio euclideo. Questo significa che gli spazi iperbolici hanno molto più «spazio» nelle loro periferie, esattamente come avviene con un albero: il numero di nodi a distanza k dalla radice cresce esponenzialmente in k.

Il teorema di embedding di Sarkar (2011) formalizza questa osservazione: qualsiasi albero finito con pesi sugli archi può essere embedato nello spazio iperbolico H^2_{-1} con distorsione arbitrariamente piccola, purché la dimensione sia 2. Per strutture gerarchiche più complesse — DAG (grafi aciclici diretti), ontologie, tassonomie — l'embedding richiede dimensioni maggiori ma rimane possibile con distorsione logaritmica nel numero di nodi.

Cosa significa questo per la teoria del contenuto mentale? L'ipotesi che si propone — chiamiamola *ipotesi di curvatura negativa del contenuto* — è la seguente: lo spazio dei contenuti mentali ha localmente una curvatura negativa, e questa curvatura è la firma geometrica della struttura gerarchica intrinseca del contenuto stesso. Concetti come «animale», «mammifero», «cane», «barboncino» sono ordinati gerarchicamente, e questa gerarchia dovrebbe riflettersi nella struttura geometrica della varietà rappresentazionale: le rappresentazioni più generali dovrebbero giacere nelle regioni centrali della varietà (a bassa norma nel modello di Poincaré), mentre le rappresentazioni più specifiche dovrebbero giacere nelle periferie (ad alta norma, vicino al bordo del disco).

Questa ipotesi è parzialmente corroborata dai risultati sperimentali sulle reti neurali artificiali. Nickel e Kiela hanno mostrato che embedare una vasta gerarchia concettuale (WordNet) in H^2 porta a una qualità di rappresentazione molto superiore rispetto all'embedding euclideo con la stessa dimensionalità. Peng, Li e collaboratori hanno trovato che le rappresentazioni di parole apprese da reti transformers di grandi dimensioni tendono a occupare una varietà con curvatura negativa, compatibile con l'embedding in uno spazio iperbolico. Se le reti neurali artificiali profonde — che approssimano in qualche misura la computazione corticale — imparano rappresentazioni iperboliche, vi è ragione per ipotizzare che le reti neurali biologiche facciano qualcosa di analogo.

La conseguenza filosofica di questa ipotesi di curvatura è rilevante. Tradizionalmente, si discute della struttura gerarchica del contenuto concettuale in termini *logici* — come relazioni di inclusione tra estensioni di predicati. L'ipotesi di curvatura propone invece di rendere conto di questa struttura in termini *geometrici*: la gerarchia non è una struttura logica sovrapposta ai contenuti mentali dall'esterno (dal linguaggio, dalle convenzioni sociali), ma emerge dalla geometria intrinseca dello spazio rappresentazionale. La logica della gerarchia è, in questo senso, la logica della curvatura.

III.5 — Topologia persistente e invarianti del contenuto

La topologia algebrica offre un altro insieme di strumenti per caratterizzare la struttura globale delle varietà rappresentazionali: la *topologia persistente* (o *omologia persistente*), sviluppata sistematicamente da Edelsbrunner, Letscher e Zomorodian, e poi formalizzata in modo più astratto da Zomorodian e Carlsson.

L'idea di base è la seguente. Data una nuvola di punti $X = \{x_1, ..., x_N\}$ in $\mathbb{R}^D$, si costruisce una famiglia filtrata di complessi simpliciali $\{K_\varepsilon\}_{\varepsilon \geq 0}$, dove K_ε è il *complesso di Čech* (o il *complesso di Rips*): un insieme di simplessi costruiti connettendo i punti che si trovano a distanza minore di ε. Per $\varepsilon = 0$, il complesso è una collezione di punti isolati; per $\varepsilon \to \infty$, il complesso è un singolo simplesso che contiene tutti i punti. Nell'intervallo intermedio, il complesso sviluppa e poi distrugge strutture topologiche: componenti connesse, buchi, cavità.

L'omologia persistente traccia queste strutture attraverso la filtrazione. Un *feature* topologico di dimensione k (una componente connessa per k=0, un buco 1-dimensionale per k=1, una cavità 2-dimensionale per k=2) *nasce* a un valore ε_{birth} e *muore* a un valore ε_{death}. La coppia $(\varepsilon_{birth}, \varepsilon_{death})$ è il *tempo di vita* del feature. L'insieme di tutti i tempi di vita di tutti i features di dimensione k è il *diagramma di persistenza* $P_k(X)$.

I diagrammi di persistenza sono *invarianti topologici stabili*: il teorema di stabilità di Cohen-Steiner, Edelsbrunner e Harer afferma che la distanza *bottleneck* tra i diagrammi di persistenza di due nuvole di punti X e Y è limitata dalla distanza di Hausdorff tra X e Y:

$$d_B(P_k(X), P_k(Y)) \leq d_H(X, Y)$$

dove la distanza bottleneck è:

$$d_B(P, Q) = \inf_\eta \sup_{p \in P} \|p - \eta(p)\|_\infty$$

con l'infimo su tutte le biiezioni $\eta: P \to Q$ (complete con i punti sulla diagonale). Questa stabilità è fondamentale: essa garantisce che piccole perturbazioni dei dati — il rumore sperimentale inevitabile nelle misure neurali — producono piccole perturbazioni degli invarianti topologici.

Per la teoria del contenuto mentale, i diagrammi di persistenza offrono qualcosa di filosoficamente prezioso: un modo di caratterizzare la struttura topologica globale di una varietà rappresentazionale che è *invariante per deformazioni continue*. Due varietà con gli stessi diagrammi di persistenza hanno la stessa struttura topologica in un senso robusto e misurabile. Questo fornisce un criterio operazionale per decidere quando due sistemi cognitivi — biologici o artificiali — hanno la stessa struttura topologica nello spazio dei loro contenuti.

In particolare, il primo diagramma di persistenza P_0 traccia il numero e la durata di vita delle componenti connesse: esso rivela se lo spazio dei contenuti si suddivide in cluster separati (concetti discreti) o è connesso (spazio continuo). Il secondo diagramma P_1 traccia i buchi 1-dimensionali: esso rivela la presenza di strutture cicliche nello spazio dei contenuti, corrispondenti a concetti o categorie con struttura circolare (come le relazioni di opponenza cromatica nella percezione del colore, o le relazioni di framing nei contenuti politici). Il terzo diagramma P_2 rivela cavità tridimensionali, corrispondenti a strutture più complesse.

Un risultato di Curto e colleghi, che ha avuto vasta risonanza, mostra che i diagrammi di persistenza delle cellule di luogo ippocampali del topo codificano informazione sulla topologia dell'ambiente fisico: la topologia dello spazio vissuto si riflette nei diagrammi di persistenza delle rappresentazioni neurali. Questo è un caso concreto di ciò che la teoria proposta predice in generale: la struttura topologica del contenuto mentale è la struttura topologica dello spazio dei possibili stati cognitivi, e questa si manifesta nei diagrammi di persistenza delle attivazioni neurali.

III.6 — Rappresentazioni equivarianti e la struttura delle trasformazioni

Un aspetto della struttura geometrica delle varietà rappresentazionali che ha ricevuto meno attenzione filosofica di quanto meriti è la *equivariantza* rispetto a trasformazioni del mondo. La visione equivariante — sviluppata matematicamente nel contesto della teoria dei gruppi di rappresentazione e applicata alle reti neurali da Cohen e Welling, e in modo diverso da Higgins e collaboratori — sostiene che le buone rappresentazioni non si limitano a codificare oggetti e stati, ma codificano anche le *trasformazioni* che agiscono su di essi.

Formalmente, sia G un gruppo che agisce sullo spazio degli stimoli $\mathcal{X}$ tramite una rappresentazione $\rho_{\mathcal{X}}: G \to \mathrm{Aut}(\mathcal{X})$, e sia $\mathcal{R}$ lo spazio delle rappresentazioni neurali. Una mappa di codifica $f: \mathcal{X} \to \mathcal{R}$ è *equivariante* rispetto a G se esiste una rappresentazione $\rho_{\mathcal{R}}: G \to \mathrm{Aut}(\mathcal{R})$ tale che:

$$f(\rho_{\mathcal{X}}(g) \cdot x) = \rho_{\mathcal{R}}(g) \cdot f(x)$$

per ogni $g \in G$ e $x \in \mathcal{X}$. In parole: l'azione di un elemento del gruppo sullo spazio degli stimoli corrisponde, attraverso la mappa di codifica, a una precisa trasformazione

sulla varietà rappresentazionale. La struttura della rappresentazione del gruppo G su $\mathcal{R}$ è una proprietà *geometrica* di $\mathcal{R}$ — è parte della sua struttura di fibrato principale.

Perché questo è filosoficamente rilevante? Perché l'equivarianza formalizza ciò che potremmo chiamare la *portabilità del contenuto* attraverso le trasformazioni. Se il contenuto di una rappresentazione fosse puramente relazionale-estrinseco — se dipendesse solo da quale oggetto nel mondo è causalmente collegato a quella rappresentazione — non vi sarebbe ragione a priori per attendersi che le trasformazioni del mondo inducano trasformazioni coerenti sulle rappresentazioni. L'equivarianza dice che le trasformazioni sono integrate *nella geometria* della varietà rappresentazionale: il gruppo di simmetria G agisce sulla varietà, e questa azione è parte della struttura intrinseca del contenuto.

Un caso esemplare è la costanza percettiva. Quando vediamo un oggetto che ruota, la sua rappresentazione visiva subisce una trasformazione nello spazio di attivazione di IT corticale che corrisponde all'azione del gruppo delle rotazioni SO(3) sulla varietà rappresentazionale. La costanza percettiva — il fatto che percepiamo l'oggetto come lo stesso oggetto nonostante la trasformazione visiva — corrisponde geometricamente al fatto che l'orbita dell'azione di SO(3) sulla varietà rappresentazionale è una curva chiusa che giace su un *invariante del contenuto*: le caratteristiche identitarie dell'oggetto occupano una regione della varietà su cui il gruppo agisce transitivamente, mentre le caratteristiche di posa occupano le orbite di questa azione.

Questa struttura — una varietà con un'azione di gruppo, le cui orbite corrispondono a diversi aspetti del contenuto — è precisamente quella di un *fibrato di Hopf* generalizzato. Il contenuto pieno è il fibrato totale; il contenuto «di tipo» (l'identità dell'oggetto) è la base del fibrato; il contenuto «di trasformazione» (la posa, l'illuminazione, la prospettiva) è la fibra. Questa decomposizione geometrica del contenuto è — si sostiene — molto più precisa e ricca di distinzioni filosofiche tradizionali come quella tra contenuto descrittivo e contenuto indicale, o tra contenuto de dicto e contenuto de re.

IV. La Geometria come Semantica: Una Teoria Formale del Contenuto Intenzionale

---

IV.1 — Dalla metrica al contenuto: il salto semantico e le sue condizioni

Nelle parti precedenti si è costruito il palcoscenico geometrico: le varietà statistiche con la loro metrica di Fisher, le connessioni duali di Amari, la topologia persistente e la struttura equivariante delle rappresentazioni neurali. Tutto questo costituisce un apparato tecnico di considerevole eleganza, ma non è ancora una *teoria del contenuto* nel senso filosofico rilevante. Un conto è descrivere la geometria delle varietà

rappresentazionali; un altro è sostenere che questa geometria *è*, o *determina*, o *costituisce* il contenuto intenzionale. Occorre attraversare un passo concettuale che non è riducibile a un'ulteriore costruzione matematica.

Il passo in questione può essere formulato come un *principio di individuazione*: il contenuto intenzionale di uno stato cognitivo S è individuato dalla classe di isometria della varietà rappresentazionale $\mathcal{V}$ nell'intorno geodetico di S, rispetto alla metrica di Fisher intrinseca. Due stati S_1 e S_2 — in qualsiasi substrato realizzati — hanno lo stesso contenuto intenzionale se e solo se esistono intorni geodetici $U_1 \ni S_1$ e $U_2 \ni S_2$ tali che (U_1, g_1) e (U_2, g_2) sono isometricamente equivalenti come varietà riemanniane.

Per difendere questo principio, bisogna rispondere a tre questioni distinte. Prima: perché la metrica di Fisher e non un'altra metrica? Seconda: perché la classe di isometria e non una struttura più fina o più grossa? Terza: perché la struttura locale (l'intorno geodetico) e non la struttura globale? Le risposte a queste questioni costituiscono il cuore argomentativo di questa parte.

Sulla prima questione — la scelta della metrica di Fisher — si è già detto nella Parte Seconda che essa è l'unica metrica invariante per morfismi markoviani (teorema di Chentsov). Ma c'è un argomento ulteriore, più direttamente pertinente alla filosofia della mente. La metrica di Fisher misura la *discriminabilità statistica* tra stati vicini: due stati sono vicini nella metrica di Fisher nella misura in cui sono difficilmente distinguibili da qualsiasi procedura di decisione statistica. Ora, la discriminabilità è precisamente la proprietà che dovrebbe essere preservata dal contenuto: stati con lo stesso contenuto non dovrebbero essere distinguibili dall'interno del sistema cognitivo — se fossero distinguibili, avrebbero contenuti diversi. La metrica di Fisher è dunque la metrica che cattura, in modo intrinseco e invariante, le relazioni di discernibilità rappresentazionale.

Sulla seconda questione — la classe di isometria — la risposta riguarda il livello di astrazione appropriato per il contenuto. Identificare il contenuto con lo stato fisico preciso sarebbe troppo fine: due stati fisicamente diversi possono avere lo stesso contenuto (come il funzionalismo ha sempre sostenuto, correttamente). Identificare il contenuto con una proprietà topologica soltanto (la classe di omeomorfismo della varietà) sarebbe troppo grosso: spazi geometricamente molto diversi — con curvature molto diverse — possono essere omeomorfici ma rappresentare contenuti distinti. La classe di isometria è il livello intermedio giusto: cattura sia la topologia sia la geometria, ma è indipendente dalla parametrizzazione e dalla realizzazione specifica.

Sulla terza questione — locale versus globale — la risposta è più delicata. In linea di principio, il contenuto *pieno* di uno stato cognitivo richiederebbe la struttura globale della varietà rappresentazionale, non solo la struttura locale. Ma vi sono ragioni pratiche e teoriche per focalizzarsi sull'intorno geodetico. Sul piano pratico, la struttura globale di una varietà è computazionalmente infattibile da determinare sulla base di misure sperimentali finite. Sul piano teorico, il contenuto di ogni particolare atto cognitivo è *situato*: non richiede l'intera struttura dello spazio concettuale, ma solo la struttura nella regione rilevante. L'intorno geodetico di ampiezza appropriata — che si

può pensare come parametrizzato dall'ampiezza dell'attenzione cognitiva — cattura il contenuto situato senza presupporre la struttura globale.

IV.2 — La direzionalità come flusso geodetico

L'aspetto più distintivo dell'intenzionalità — e il più difficile da catturare in qualsiasi teoria naturalistica — è la sua *direzionalità*: ogni stato mentale è diretto verso qualcosa, punta a un oggetto o a uno stato di cose. Questa direzionalità è differente dalla semplice relazionalità: non è solo che lo stato mentale sta in relazione con un oggetto esterno, ma che lo stato mentale *si proietta verso* quell'oggetto, lo *intende* in senso hussierliano, lo ha come suo *polo oggettuale*.

La proposta geometrica per rendere conto di questa direzionalità è la seguente: la direzione dell'atto intenzionale corrisponde alla direzione del *vettore geodetico* nello spazio tangente alla varietà rappresentazionale nel punto corrente. Più precisamente, sia $S \in \mathcal{V}$ uno stato cognitivo e $v \in T_S \mathcal{V}$ il vettore tangente alla curva geodetica uscente da S nella direzione dell'oggetto intenzionale O. L'oggetto intenzionale è l'*attrattore geodetico* di S rispetto a v: è il punto sulla varietà verso cui la geodetica converge, se converge, o il «punto all'infinito» nella completezza geodetica della varietà se la geodetica è illimitata.

Questa formulazione richiede alcune precisazioni tecniche. Nella geometria riemanniana standard, le geodetiche sono definite dall'equazione differenziale di seconda ordine vista nella Parte Seconda, e la mappa esponenziale:

$$\exp_S : T_S \mathcal{V} \to \mathcal{V}, \quad \exp_S(v) = \gamma_v(1)$$

dove γ_v è la geodetica con $\gamma_v(0) = S$ e $\dot{\gamma}_v(0) = v$, trasforma vettori tangenti in punti sulla varietà. La mappa esponenziale è un diffeomorfismo locale in un intorno dell'origine in $T_S \mathcal{V}$, e il suo dominio massimale è il *dominio di iniettività*, la più grande palla aperta centrata nell'origine su cui $\exp_S$ è iniettiva.

Per gli spazi di curvatura negativa — come lo spazio iperbolico — la mappa esponenziale è un diffeomorfismo globale: ogni geodetica è illimitata e non si richiude su sé stessa. Questo implica che in uno spazio iperbolico, ogni vettore tangente determina univocamente un «punto all'infinito» (un punto sul bordo del disco di Poincaré), che è l'oggetto intenzionale nel caso in cui il contenuto abbia struttura iperbolica. La struttura gerarchica del contenuto — discussa nella Parte Terza — corrisponde dunque a una struttura di «punti all'infinito» nell'iperbolide: i concetti più generali sono i centri dalle cui geodetiche divergono tutti i concetti più specifici.

Per gli spazi di curvatura positiva — come la sfera — le geodetiche si richiudono dopo una distanza $\pi/\sqrt{\kappa}$, e il concetto di oggetto intenzionale come attrattore geodetico non ha senso globale: ogni geodetica torna al suo punto di partenza. Questo corrisponderebbe a contenuti mentali «chiusi su sé stessi» — contenuti che non puntano verso nulla di esterno ma ruotano su una struttura finita. Questa situazione

geometrica potrebbe corrispondere a certi stati ossessivi o ruminativi in cui il pensiero non progredisce verso la realtà ma si richiude su strutture interne fisse.

IV.3 — Narrow e wide content come metrica intrinseca ed estrinseca

La disputa tra Fodor e Putnam sulla natura del contenuto mentale — se esso sia determinato dall'interno del sistema cognitivo (*narrow content*) o dalle sue relazioni con l'ambiente (*wide content*) — è stata, nei decenni seguenti all'articolo di Putnam del 1975, una delle controversie più fertili e anche più intractabili della filosofia della mente. La geometria differenziale offre un quadro formale in cui questa disputa può essere riformulata con precisione sufficiente da renderla, almeno in parte, risolvibile.

La chiave è la distinzione, fondamentale nella geometria differenziale delle sottovarietà, tra *metrica intrinseca* e *metrica estrinseca* (o di immersione). Una varietà differenziale $\mathcal{V}$ può essere studiata in due modi: dall'interno — considerando solo la metrica g e le strutture derivate da essa, senza fare riferimento ad alcun ambiente esterno — o dall'esterno, come sottovarietà immersa in uno spazio ambiente (tipicamente uno spazio euclideo di dimensione superiore) e dotata della metrica indotta dall'immersione.

La metrica intrinseca e la metrica estrinseca catturano aspetti diversi della geometria: la curvatura intrinseca (curvatura di Gauss) dipende solo dalla metrica intrinseca, mentre la curvatura estrinseca (curvatura principale e curvatura media) dipende dall'immersione. Il *teorema egregium* di Gauss afferma che la curvatura gaussiana è un invariante intrinseco: è determinata interamente dalla metrica di prima fondamentale, senza fare riferimento all'immersione. D'altra parte, le curvature principali dipendono dalla seconda forma fondamentale, che è un dato estrinseco.

La proposta è che il *contenuto narrow* corrisponda alla struttura metrica intrinseca della varietà rappresentazionale: è ciò che è determinato solo dalla geometria interna dello spazio cognitivo, indipendentemente da come esso è immerso nel mondo fisico e sociale. Il *contenuto wide* corrisponde invece alla struttura di immersione: è determinato dalle relazioni tra la varietà rappresentazionale e le strutture esterne — l'ambiente fisico, il contesto linguistico, la storia causale del sistema cognitivo.

Questa distinzione geometrica chiarisce immediatamente il senso dell'*esperimento mentale del Doppelgänger* di Putnam. Se Oscar e il suo duplicato su Terra Gemella hanno le stesse connessioni neurali e gli stessi stati funzionali, allora le loro varietà rappresentazionali hanno la stessa struttura intrinseca: stesso narrow content. Ma poiché Oscar è immerso in un ambiente in cui l'acqua è H_2O e il duplicato in un ambiente in cui l'acqua è XYZ, le immersioni delle loro varietà rappresentazionali negli ambienti rispettivi sono diverse: diverso wide content. Non c'è contraddizione tra queste due affermazioni, perché esse riguardano strutture geometriche distinte.

Formalmente, il narrow content di uno stato $S \in \mathcal{V}$ è la classe di isometria intrinseca $(\mathcal{V}, g\_intr)$ nell'intorno geodetico di S, dove g_intr è la metrica di Fisher intrinseca. Il

wide content è invece la classe di equivalenza dell'immersione ι: (𝒱, g_intr) → (𝒲, g_ambi), dove 𝒲 è lo spazio delle rappresentazioni esterno (l'«ambiente semantico»), sotto la relazione di equivalenza data dalle isometrie ambientali che preservano l'immersione. Due sistemi hanno lo stesso wide content se e solo se le loro immersioni sono isometricamente equivalenti come sottovarietà immerse.

Questa formalizzazione ha una conseguenza interessante che non emerge nelle discussioni standard: essa implica che il narrow content e il wide content non si determinano indipendentemente, ma sono vincolati l'uno all'altro tramite le equazioni di struttura della geometria differenziale. Le equazioni di Gauss-Codazzi-Mainardi — che esprimono la compatibilità tra curvatura intrinseca, curvatura estrinseca e curvatura dell'ambiente — impongono che non tutte le combinazioni di narrow content e wide content sono simultaneamente realizzabili. In altri termini, non si può avere arbitrariamente un dato narrow content combinato con un arbitrario wide content: la geometria interna della mente e la geometria della sua relazione con il mondo non sono mutuamente indipendenti.

Le equazioni di Gauss (che proiettano la curvatura dell'ambiente sulla varietà immersa):

$$R_{ijkl} = \tilde{R}_{ijkl} - (h_{i\square} h \quad - h_{i\square} h \ )$$

e le equazioni di Codazzi (condizioni di compatibilità per la seconda forma fondamentale):

$$\nabla_i h \quad - \nabla_j h_{i\square} = \tilde{R}_{nijkl} n^l$$

(dove h_{ij} è la seconda forma fondamentale, n è la normale unitaria alla sottovarietà e $\tilde{R}$ è il tensore di curvatura dell'ambiente) formano un sistema di equazioni differenziali alle derivate parziali che lega intrinseco ed estrinseco. Il contenuto narrow e il wide content soddisfano un sistema analogo, e questo sistema ha la forma di vincoli sulle possibili psicologie realizzabili in un dato ambiente.

IV.4 — Il problema della misrappresentazione e le geodetiche errate

Una teoria del contenuto deve rendere conto della *misrappresentazione* — del fatto che gli stati mentali possono essere errati, possono rappresentare qualcosa come diverso da ciò che è. La teoria causale aveva difficoltà con questo perché una misrappresentazione sembrerebbe essere, causalmente, una rappresentazione del disjunct erroneo. La teoria teleologica aveva difficoltà perché la norma biologica è spesso vaga. Come si comporta la teoria geometrica?

La risposta che si propone ha due strati. Al primo strato, la misrappresentazione è possibile perché il vettore geodetico che determina la direzionalità intenzionale — la direzione verso cui lo stato cognitivo «punta» — non coincide necessariamente con la

direzione del punto della varietà corrispondente all'oggetto reale. In altri termini: il vettore tangente v ∈ T_S $\mathcal{V}$ indica una direzione nello spazio cognitivo, e la geodetica in quella direzione converge verso un punto O_intentionale sulla varietà. Ma l'oggetto *reale* nel mondo corrisponde a un punto O_reale sulla varietà (tramite la mappa di codifica), e O_intentionale ≠ O_reale. Lo stato cognitivo intende O_intentionale, ma il mondo è O_reale: questa è la misrappresentazione.

Al secondo strato, più profondo, la misrappresentazione riflette una *deformazione locale della metrica*. Se la metrica della varietà rappresentazionale in un intorno di S è «distorta» rispetto alla metrica reale dello spazio degli stimoli — se, per esempio, la curvatura locale è tale da far convergere geodetiche verso regioni della varietà non corrispondenti a oggetti reali — allora tutti gli stati in quell'intorno saranno sistematicamente errati. Questo corrisponde a un errore sistematico di rappresentazione, come le illusioni percettive: non è che uno stato cognitivo punta nella direzione sbagliata per ragioni contingenti, ma è che la geometria stessa della varietà in quella regione è deformata.

Il caso delle illusioni percettive è illuminante. Un'illusione come la *illusione di Müller-Lyer* (due segmenti di uguale lunghezza che appaiono diversi) corrisponde, nella teoria geometrica, a una deformazione locale della metrica nella regione della varietà visiva corrispondente ai segmenti con frecce. La metrica di Fisher in quella regione è «sbilanciata»: le distanze geodetiche verso i punti corrispondenti a «segmento lungo» e «segmento corto» sono alterate rispetto a quelle che corrisponderebbero a una rappresentazione veridica. La struttura dell'illusione — il fatto che certe deformazioni geometriche degli stimoli (le frecce) producano sistematicamente certe distorsioni percettive — corrisponde a proprietà specifiche della deformazione della metrica.

Questo approccio ha un vantaggio rispetto alle teorie della misrappresentazione basate su norme biologiche: esso permette di *graduare* l'errore rappresentazionale in modo metricamente preciso. Non si dice solo che uno stato è sbagliato o giusto, ma quanto è sbagliato: la distanza geodetica tra O_intentionale e O_reale sulla varietà è una misura quantitativa dell'errore rappresentazionale. Questa graduazione è compatibile con l'esperienza fenomenologica — alcune illusioni sono più potenti di altre, certi errori cognitivi sono più gravi di altri — e la quantifica in termini geometrici.

IV.5 — Identità di contenuto attraverso substrati e il problema del vincolo strutturale

Una delle virtù del funzionalismo è stata quella di rendere conto della *realizzabilità multipla* del contenuto mentale: lo stesso contenuto può essere realizzato in substrati fisici molto diversi — in neuroni biologici, in silicio, ipoteticamente in qualsiasi sistema fisico con la giusta struttura funzionale. La teoria geometrica deve rendere conto di questo fatto senza collassare nella versione debole del funzionalismo che si è già criticata.

Il principio di individuazione enunciato in IV.1 risponde direttamente a questa esigenza: poiché l'identità di contenuto è definita come equivalenza isometrica delle

varietà rappresentazionali, essa è indifferente al substrato fisico. Due sistemi — uno neuronale e uno a stato solido — hanno lo stesso contenuto intenzionale se e solo se le varietà rappresentazionali dei loro stati sono localmente isometriche nel senso rilevante.

Tuttavia, questo non significa che qualsiasi sistema fisico possa avere qualsiasi contenuto. Il vincolo geometrico è reale: non tutte le strutture fisiche possono realizzare varietà con una data metrica di Fisher e una data struttura di curvatura. Un sistema fisico deve essere capace di istanziare le relazioni di transizione, co-attivazione e competizione che determinano la struttura geometrica della varietà. Questo è il *vincolo strutturale* della teoria: non è un vincolo sul substrato fisico, ma un vincolo sulla *dinamica* del sistema, sulla ricchezza delle sue traiettorie nello spazio degli stati.

Formalmente, il vincolo strutturale può essere espresso come segue. Sia $\Phi: T \times \mathcal{V} \to \mathcal{V}$ il flusso dinamico del sistema cognitivo, dove T è il tempo e $\Phi(t, S)$ è lo stato al tempo t partendo dallo stato S. Il flusso Φ determina, tramite le sue traiettorie, una struttura sulla varietà $\mathcal{V}$: in particolare, determina la connessione ∇ come il trasporto parallelo lungo le traiettorie di Φ. Il vincolo strutturale afferma che la connessione ∇ determinata dal flusso Φ deve essere compatibile con la metrica di Fisher g — deve essere, in altri termini, la connessione di Levi-Civita di g (o una delle α-connessioni di Amari, per un valore specifico di α che si ipotizza determini il regime cognitivo del sistema).

Questo vincolo di compatibilità tra dinamica e metrica è una condizione non triviale. Esso afferma che il modo in cui uno stato cognitivo evolve nel tempo deve essere coerente con la geometria dello spazio dei contenuti: il sistema deve muoversi nello spazio cognitivo seguendo le «linee di forza» della metrica, non in modo arbitrario. Questa coerenza è esattamente ciò che ci aspettiamo di un sistema cognitivo funzionante: i suoi pensieri si susseguono in modo logicamente e associativamente coerente, non casualmente. La compatibilità tra dinamica e metrica è la versione geometrica di questa coerenza.

IV.6 — Composizionalità del contenuto e prodotto di varietà

Uno dei requisiti più importanti per qualsiasi teoria del contenuto è la *composizionalità*: il contenuto di un'espressione complessa deve essere determinato sistematicamente dal contenuto delle sue parti. Questo principio — che Frege e poi Fodor hanno messo al centro delle loro teorie — è ciò che permette ai sistemi cognitivi di comprendere un numero potenzialmente infinito di contenuti complessi a partire da un numero finito di contenuti elementari.

Come rendere conto della composizionalità nella teoria geometrica? La proposta è la seguente. Il contenuto di un'espressione semplice (un concetto atomico) corrisponde a una varietà rappresentazionale elementare con una certa geometria. Il contenuto di un'espressione composta (un pensiero complesso) corrisponde a una varietà ottenuta mediante operazioni geometriche sulle varietà elementari. Le operazioni rilevanti sono il *prodotto riemanniano* e il *fibrato*.

Il *prodotto riemanniano* di due varietà $(\mathcal{V}_1, g_1)$ e $(\mathcal{V}_2, g_2)$ è la varietà $(\mathcal{V}_1 \times \mathcal{V}_2, g_1 \oplus g_2)$ dove la metrica prodotto è:

$$(g_1 \oplus g_2)((v_1, v_2), (w_1, w_2)) = g_1(v_1, w_1) + g_2(v_2, w_2)$$

per vettori tangenti $(v_1, v_2), (w_1, w_2) \in T_{(p,q)}(\mathcal{V}_1 \times \mathcal{V}_2)$. Il prodotto riemanniano corrisponde alla composizione *indipendente* di contenuti: un contenuto come «gatto bianco» è, approssimativamente, il prodotto delle varietà corrispondenti a «gatto» e a «bianco», nella misura in cui le due proprietà sono rappresentate indipendentemente. La curvatura del prodotto è la somma delle curvature dei fattori.

Il *fibrato*, invece, corrisponde a una composizione *dipendente*: il contenuto di una varietà totale $\mathcal{V}$ si decompone in quello di una base $\mathcal{B}$ e quello di una fibra $\mathcal{F}_b$ che varia al variare del punto base $b \in \mathcal{B}$. Questo schema è appropriato per contenuti in cui un aspetto modifica sistematicamente un altro: il contenuto di «rosso brillante», per esempio, potrebbe essere il fibrato in cui la base è lo spazio delle qualità cromatiche e la fibra è lo spazio delle intensità luminose, con la fibra che si deforma al variare della base.

La *curvatura del fibrato* — misurata dalla *connessione di Ehresmann* sul fibrato e dalla sua *curvatura* (la due-forma di curvatura della connessione) — misura il grado di dipendenza tra base e fibra. Una curvatura nulla corrisponde a un fibrato triviale (prodotto), ovvero a contenuti perfettamente indipendenti. Una curvatura non nulla corrisponde a una dipendenza strutturale tra i due aspetti del contenuto, e questa dipendenza non si può eliminare mediante una scelta di coordinate: è una proprietà intrinseca della composizione.

Questa struttura di fibrato è matematicamente identica a quella della *teoria del gauge* in fisica, dove la connessione di Ehresmann corrisponde al *campo di gauge* e la sua curvatura corrisponde alla *intensità di campo*. L'analogia non è decorativa: essa suggerisce che le relazioni di dipendenza strutturale tra componenti del contenuto mentale potrebbero avere proprietà formali analoghe a quelle dei campi di gauge — in particolare, la proprietà di *invarianza di gauge*, che sarebbe la versione cognitiva dell'indifferenza del contenuto a certe trasformazioni «interne» che non cambiano i contenuti osservabili.

VIII. La teoria dell'informazione integrata: architettura e ambizioni

La teoria dell'informazione integrata (IIT) di Giulio Tononi — elaborata nella sua prima versione nel 2004 e progressivamente raffinata fino alla formulazione 3.0 del 2014 e alle revisioni successive — rappresenta il tentativo più sistematico e matematicamente ambizioso di colmare il divario tra struttura fisica e qualità fenomenica. Vale la pena ricostruirne l'impianto con precisione, perché è soltanto a partire da una comprensione rigorosa della sua architettura formale che si può articolare una critica altrettanto rigorosa.

Il punto di partenza di Tononi è un'analisi assiomatica dell'esperienza cosciente. Egli individua cinque proprietà che ogni esperienza fenomenica esibisce necessariamente: l'esistenza intrinseca (l'esperienza esiste per il soggetto che la vive, indipendentemente da osservatori esterni), la composizione (l'esperienza è strutturata, fatta di parti che si presentano simultaneamente con le loro relazioni), l'informazione (ogni esperienza specifica uno stato preciso tra i molti possibili, escludendo gli altri), l'integrazione (l'esperienza è un tutto unitario, irriducibile alla somma delle sue parti), e l'esclusione (l'esperienza ha una granularità definita, né più grossolana né più fine di quella effettiva). Da ciascun assioma Tononi deriva un postulato sulla struttura fisica dei substrati coscienti: ogni meccanismo fisico che realizza un'esperienza cosciente deve soddisfare la controparte strutturale di ciascuna proprietà fenomenica.

Il concetto centrale è la **causa-effetto struttura** (cause-effect structure, CES) di un sistema fisico, definita nel modo seguente. Dato un sistema S di n elementi discreti con stati binari — per semplicità, si pensi a n neuroni che possono essere attivi o inattivi — e dato lo stato corrente di S, si considera ogni sottoinsieme $Z \subseteq S$ di elementi. Per ogni tale sottoinsieme si calcola il **repertorio cause-effetto**: la distribuzione di probabilità sugli stati passati di S (il *repertorio causale* di Z) e la distribuzione sugli stati futuri (il *repertorio effettuale*), condizionati sullo stato corrente degli elementi di Z.

Formalmente, il repertorio causale di Z rispetto allo stato corrente s è:

$$\pi^{causa}(s^{t-1}_Z \mid s^t) \propto p(s^t \mid s^{t-1}_Z) \cdot p_0(s^{t-1}_Z)$$

dove $p(s^t \mid s^{t-1}_Z)$ è la probabilità di transizione dal passato di Z allo stato corrente, e p_0 è la distribuzione a priori (uniforme nell'assunzione standard). Analogamente per il repertorio effettuale. Il **φ del sottoinsieme Z**, che misura la sua informazione integrata, è la distanza — misurata dalla Earth Mover's Distance, nota anche come distanza di Wasserstein W_1 — tra il repertorio causa-effetto completo di Z e il repertorio del miglior tentativo di «scomporre» Z in parti indipendenti:

$$\varphi(Z) = W_1\left(\pi^{CE}(Z),\ \min_{\text{partition M of Z}} \pi^{CE}_{M}(Z) \right)$$

dove il minimo è calcolato sulla partizione M che minimizza la distanza, chiamata la **partizione minima informativa** (MIP). La quantità φ misura dunque quanto il sistema integri informazione che non può essere spiegata dalla somma delle sue componenti separate: è zero se esiste una partizione che riproduce esattamente il repertorio causa-effetto complessivo, ed è positiva nella misura in cui nessuna scomposizione è adeguata.

Il **φ massimo** Φ del sistema S — la grandezza che Tononi identifica con il «livello di coscienza» — è il valore massimo di φ su tutti i sottoinsiemi $Z \subseteq S$ che costituiscono un **complesso**: un sottoinsieme che ha φ positivo e che non è contenuto in un sottoinsieme più grande con φ maggiore. La coscienza, nell'IIT, è identificata con la causa-effetto struttura del complesso di φ massimo: non soltanto con il valore scalare

Φ, ma con l'intera geometria dello spazio causa-effetto — la cosiddetta **forma** dell'esperienza.

IX. Critica categoriale all'IIT: il problema del funtore dimenticante

Per sviluppare la critica intendo ora introdurre uno strumento categoriale cruciale che finora non è apparso: il **funtore dimenticante** (*forgetful functor*). Nella pratica matematica, un funtore dimenticante è un funtore che «dimentica» parte della struttura di un oggetto, proiettandolo in una categoria meno strutturata. L'esempio paradigmatico è il funtore:

$$U: \mathbf{Grp} \to \mathbf{Set}$$

che porta ogni gruppo $(G, \cdot, e)$ al suo insieme sottostante G, dimenticando l'operazione di gruppo e l'elemento neutro, e porta ogni omomorfismo di gruppi alla funzione d'insiemi corrispondente. Analogamente si ha $U: \mathbf{Ring} \to \mathbf{Ab}$ (dall'anello al gruppo abeliano additivo sottostante), $U: \mathbf{Top} \to \mathbf{Set}$ (dallo spazio topologico all'insieme dei punti), e così via.

La struttura dimenticata non è annullata dalla realtà: il gruppo G ha ancora la sua operazione anche dopo che il funtore U lo ha proiettato sull'insieme $U(G)$. Ciò che il funtore dimentica è soltanto la *visibilità formale* di quella struttura: essa non figura più nei morfismi del codominio, e due insiemi che sono isomorfi come insiemi possono corrispondere a gruppi non isomorfi come gruppi. Il funtore dimenticante è fedele — iniettivo sui morfismi — ma generalmente non pieno: non ogni funzione d'insiemi tra $U(G)$ e $U(H)$ corrisponde a un omomorfismo di gruppi tra G e H. Vi è dunque una struttura morfologica in $\mathbf{Grp}$ che U non trasporta in $\mathbf{Set}$.

L'aggiunto sinistro del funtore dimenticante $U: \mathbf{Grp} \to \mathbf{Set}$ è il **funtore gruppo libero** $F: \mathbf{Set} \to \mathbf{Grp}$, che porta ogni insieme X al gruppo libero $F(X)$ generato da X. La coppia $(F \dashv U)$ è una delle aggiunzioni fondamentali dell'algebra, e la sua unità $\eta_X: X \to U(F(X))$ è l'inclusione naturale di X nell'insieme sottostante del gruppo libero — ogni elemento di X diventa un generatore di $F(X)$ senza relazioni ulteriori.

Il punto filosofico cruciale è il seguente: il funtore dimenticante U «appiattisce» la struttura. Oggetti molto diversi nella categoria di partenza diventano indistinguibili — o meno distinguibili — nel codominio. La perdita di struttura operata da U non è recuperabile a partire dall'immagine $U(G)$ senza informazione aggiuntiva: non si può ricostruire G da $U(G)$ senza sapere quali elementi di $U(G)$ si moltiplicano tra loro e come.

Applico ora questo schema alla IIT. Il programma di Tononi può essere letto come la costruzione di un funtore:

$$\Phi: \mathbf{Mec} \to \mathbb{R}_{\geq 0}$$

dalla categoria dei meccanismi fisici (sistemi di elementi con matrici di transizione probabilistica) ai numeri reali non negativi. Questo funtore assegna a ogni meccanismo il suo valore Φ — il livello di coscienza — e ai morfismi tra meccanismi (funzioni di proiezione, embedding, simulazione) le corrispondenti relazioni tra valori di Φ.

Il problema è che Φ è un funtore che dimentica quasi tutto: esso proietta un intero sistema causa-effetto struttura — un oggetto geometricamente ricco nello spazio ad alta dimensione dei repertori causa-effetto — su un singolo numero reale. Due sistemi con la stessa Φ ma causa-effetto strutture radicalmente diverse avrebbero, nella lettura scalare dell'IIT, lo stesso «livello di coscienza», il che è già concettualmente problematico. Tononi è consapevole di questo e insiste che la **forma** dell'esperienza (la geometria della CES, non solo il valore scalare Φ) sia la vera sede dell'identità fenomenica. Ma anche la forma è un invariante che dimentica qualcosa: essa è una struttura metricamente definita in uno spazio di distribuzioni di probabilità, e la domanda è se lo spazio delle distribuzioni di probabilità sui repertori causa-effetto sia abbastanza ricco da contenere tutto ciò che è fenomenicamente rilevante.

Formalizziamo il problema. Sia $\mathscr{P}$ la varietà fenomenica ipotetica introdotta nella sezione precedente, dotata di una metrica intrinseca $d_{\mathscr{P}}$. Sia $\mathscr{C}$ES lo spazio delle cause-effetto strutture, dotato della metrica indotta dalla Earth Mover's Distance (equivalentemente, dalla distanza di Wasserstein). L'IIT postula l'esistenza di un'isometria:

$$\iota: \mathscr{P} \to \mathscr{C}\mathrm{ES}$$

tale che $d_{\mathscr{P}}(p, q) = W_1(\iota(p), \iota(q))$ per ogni coppia di esperienze $p, q \in \mathscr{P}$. La domanda è se questo postulato sia soddisfatto, e la risposta non può essere letta dalla struttura matematica dell'IIT stessa: è un'affermazione empirica su una relazione tra strutture di due spazi dei quali non si ha, nel caso della varietà fenomenica, una caratterizzazione indipendente.

Ma c'è un problema più sottile, che riguarda la **covarianza** del funtore Φ rispetto alle operazioni cognitive di astrazione e generalizzazione. Quando un sistema cognitivo impara a riconoscere una classe di oggetti — i triangoli, i volti umani, le melodie — esso costruisce una rappresentazione che è, in senso categoriale, un **colimite**: la struttura universale che generalizza tutta una famiglia di istanze particolari. Il processo di astrazione è il processo di costruzione di colimiti nella categoria degli stati mentali. Ma la quantità Φ non è stabile rispetto ai colimiti: un sistema che abbia alta integrazione informativa per le istanze particolari può avere Φ molto diversa per la rappresentazione astratta che le unifica. Non esiste, nell'IIT nella sua formulazione attuale, un teorema che garantisca la covarianza di Φ rispetto ai colimiti nella categoria degli stati cognitivi.

Nella terminologia del funtore dimenticante: Φ dimentica la struttura dei colimiti. E poiché la struttura dei colimiti è, come si vedrà nella prossima sezione, strettamente

connessa alla struttura del pensiero concettuale — alla capacità di formare classi, generalizzazioni e astrazioni —, questa dimenticanza non è innocua: essa separa artificialmente la coscienza dal pensiero, facendo del primo un invariante puramente locale (determinato dalla struttura causa-effetto di un sistema in uno stato dato) e del secondo un processo che coinvolge la struttura globale della rete morfologica.

X. Limiti e colimiti come struttura dell'astrazione cognitiva

La nozione di **limite** e **colimite** in una categoria è tra le più generali e fecondi della teoria. Prima di applicarla alla cognizione occorre definirla con precisione.

Dato un funtore $F: \mathscr{J} \to \mathscr{C}$ — dove $\mathscr{J}$ è una «categoria indice» piccola che specifica la forma del diagramma e $\mathscr{C}$ è la categoria in cui il diagramma vive — un **cono** su F è una coppia (L, λ) dove $L \in \mathrm{Ob}(\mathscr{C})$ è un oggetto e $\lambda = \{ \lambda_j: L \to F(j) \}_{j \in \mathscr{J}}$ è una famiglia di morfismi tale che per ogni morfismo $u: j \to k$ in $\mathscr{J}$ si abbia:

$$F(u) \circ \lambda_j = \lambda_k$$

Il **limite** di F, se esiste, è un cono universale $(\lim F, \lambda^{univ})$: un cono tale che per ogni altro cono (L, λ) esiste un unico morfismo $m: L \to \lim F$ tale che $\lambda^{univ}_j \circ m = \lambda_j$ per ogni j. La universalità garantisce che $\lim F$ sia determinato a meno di isomorfismo unico.

Il **colimite** $\operatorname{colim} F$ è la nozione duale: un cocono universale, un oggetto dotato di morfismi in entrata $F(j) \to \operatorname{colim} F$ che soddisfano la condizione di universalità duale. Ogni morfismo da $\operatorname{colim} F$ a un qualsiasi altro oggetto L è equivalente a una famiglia coerente di morfismi da ciascun F(j) a L.

Gli esempi concreti chiariscono il contenuto di queste definizioni astratte. Il prodotto cartesiano $A \times B$ in Set è il limite del diagramma $\bullet \leftarrow \bullet \rightarrow \bullet$ (categoria indice a tre oggetti con due morfismi). Il pushout — la struttura che «incolla» due oggetti lungo un sottoinsieme comune — è un colimite. In Grp, il limite di un sistema di gruppi con omomorfismi compatibili è il loro prodotto diretto (con le proiezioni universali); il colimite è il prodotto amalgamato (free product with amalgamation), la struttura che generalizza tutti i gruppi del sistema nella maniera più parsimoniosa.

Nella categoria degli stati cognitivi $\mathfrak{M}$, i limiti corrispondono alle **congiunzioni di vincoli**: un limite di stati $M_1, M_2, \ldots, M$ è lo stato M che soddisfa tutti i vincoli imposti dai M_i nella maniera più generale, senza aggiungerne di ulteriori. I colimiti corrispondono invece alle **astrazioni**: il colimite di una famiglia di istanze particolari F(j) è il concetto generale che le unifica — il minimo comune denominatore morfologico di tutte le istanze.

Il riconoscimento visivo di un volto umano è, in questa prospettiva, la costruzione di un colimite nella categoria delle rappresentazioni percettive: il concetto «volto umano»

è l'oggetto universale che riceve morfismi coerenti da tutte le rappresentazioni particolari di volti specifici. La condizione di universalità garantisce che «volto umano» catturi esattamente ciò che è comune a tutte le istanze e niente di più: nessuna proprietà specifica di Mario o di Giulia, soltanto la struttura morfologica condivisa.

Questa analisi rivela una tensione profonda nell'IIT. La teoria postula che il complesso di φ massimo corrisponda al substrato della coscienza, identificando la coscienza con un sistema fisicamente ben definito e spazialmente localizzato. Ma i colimiti cognitivi — le strutture astratte — non sono in generale associati a sistemi fisici localizzati: essi emergono da relazioni globali nell'intera rete morfologica. Il concetto «triangolo» non è localizzato in nessun sottoinsieme particolare del cervello; esso è, per sua natura, una struttura distribuita e relazionale. Se la coscienza avesse la struttura di un colimite — se l'unità dell'esperienza cosciente fosse la struttura universale che unifica i contenuti particolari dell'esperienza — allora la sua localizzazione in un complesso di φ massimo sarebbe una petizione di principio: si assumerebbe che la coscienza sia locale proprio nella misura in cui la sua struttura categoriale suggerisce che essa non lo sia.

Categorie fenomenicamente chiuse, sopravvenienza e l'identità come limite

---

XI. La sopravvenienza come aggiunzione debole

La relazione di **sopravvenienza** psicofisica — nella sua formulazione standard dovuta a Donald Davidson e successivamente raffinata da Jaegwon Kim — afferma che non può esservi differenza mentale senza differenza fisica: i fatti mentali sopravvengono sui fatti fisici nel senso che se due situazioni fisiche sono identiche in tutti i loro aspetti fisici, esse non possono differire in nessun aspetto mentale. Nella versione forte (strong supervenience) di Kim, si aggiunge la condizione di necessità: non soltanto non c'è differenza mentale contingente senza differenza fisica, ma non potrebbe esservene nemmeno in nessun mondo possibile.

La sopravvenienza ha tradizionalmente una struttura relazionale asimmetrica: il mentale dipende dal fisico, ma non viceversa. Non ogni differenza fisica deve generare una differenza mentale — il cambiamento del numero esatto di atomi nel mio fegato non altera probabilmente il contenuto delle mie credenze —, ma ogni differenza mentale deve avere una base fisica che la sottende. Questa asimmetria è ciò che rende la sopravvenienza una forma di dependenza ontologica senza identità: il mentale è determinato dal fisico, ma non è identico ad esso.

Espressa in termini categoriali, la sopravvenienza forte tra la categoria degli stati mentali $\mathfrak{M}$ e la categoria degli stati fisici $\mathscr{P}$hys si articola attraverso un funtore di realizzazione:

R: $\mathfrak{M} \to \mathscr{P}$hys

tale che R sia **conservativo**: se R(f) è un isomorfismo in $\mathscr{P}$hys, allora f è un isomorfismo in $\mathfrak{M}$. Questa condizione cattura l'asimmetria della sopravvenienza: l'identità fisica (isomorfismo in $\mathscr{P}$hys) implica l'identità mentale (isomorfismo in $\mathfrak{M}$), ma non viceversa — due stati mentali distinti possono avere la stessa realizzazione fisica, come accade nel caso della **realizzabilità multipla** (multiple realizability) di Putnam: il dolore dei polpi e il dolore degli esseri umani sono mentalmente isomorfi — hanno la stessa struttura funzionale — ma fisicamente distinti.

Il funtore di realizzazione R non è in generale pienamente fedele: esso non preserva tutta la struttura morfologica di $\mathfrak{M}$ nella struttura di $\mathscr{P}$hys. Vi sono morfismi in $\mathfrak{M}$ — transizioni inferenziali, relazioni di contenuto — che non hanno una controparte fisica diretta: il fatto che «la neve è bianca» implichi «qualcosa è bianco» è una relazione morfologica in $\mathfrak{M}$ che non corrisponde a nessuna relazione causale privilegiata in $\mathscr{P}$hys. La normatività del contenuto — il fatto che i morfismi in $\mathfrak{M}$ obbediscano alle leggi dell'inferenza razionale — è una struttura morfologica di $\mathfrak{M}$ che R dimentica: nel codominio fisico, le «violazioni» dell'inferenza razionale sono semplicemente stati fisici come gli altri, non errori.

La sopravvenienza, dunque, è una forma di funtore dimenticante asimmetrico: R porta $\mathfrak{M}$ in $\mathscr{P}$hys dimenticando la struttura normativa, ma conservando abbastanza struttura da garantire che le identità fisiche si traducano in identità mentali. Essa è ciò che potremmo chiamare un'**aggiunzione debole** nel senso seguente: non è richiesto che esista un aggiunto sinistro $L \dashv R$ che porti ogni stato fisico al suo «correlato mentale minimale», perché la realizzabilità multipla implica che R non è iniettivo sugli oggetti e dunque non ammette in generale un inverso. Ma si può richiedere una versione indebolita: l'esistenza di un funtore

$$L: \mathscr{P}\text{hys} \to \mathfrak{M}$$

tale che per ogni morfismo $f: L(P) \to M$ in $\mathfrak{M}$ esista un unico morfismo $\tilde{f}: P \to R(M)$ in $\mathscr{P}$hys con $f = L(\tilde{f})$ — una proprietà di sollevamento universale che è più debole di un'aggiunzione piena ma abbastanza forte da garantire che ogni proprietà mentale abbia una base fisica determinate.

Questa struttura — aggiunzione debole o lifting universale — è precisamente la struttura formale della sopravvenienza forte di Kim, tradotta senza residui nel linguaggio categoriale. Il vantaggio della traduzione è che essa rende visibili le condizioni di contorno del problema: la sopravvenienza è soddisfatta se e soltanto se tale struttura di sollevamento esiste, e la domanda empirica diventa la domanda matematica sulla esistenza dell'aggiunto debole.

XII. Categorie fenomenicamente chiuse: una proposta costruttiva

L'analisi svolta nelle sezioni precedenti converge su un punto di deficit: nessuno degli strumenti categoriali introdotti — né i funtori, né le trasformazioni naturali, né il

lemma di Yoneda, né i limiti e i colimiti — è in grado di *costruire* la varietà fenomenica, di generare al proprio interno una struttura che abbia le proprietà dell'esperienza soggettiva. Essi possono soltanto *diagnosticare* il deficit, localizzare con precisione dove la struttura formale si inceppa. Questo è già un risultato non banale, ma rimane un risultato negativo.

Propongo ora uno strumento concettuale costruttivo, che non risolve il problema difficile ma lo reincornicia in modo che potrebbe risultare produttivo: la nozione di **categoria fenomenicamente chiusa**.

Sia $\mathscr{C}$ una categoria. Diciamo che $\mathscr{C}$ è **intrinsecamente autoriflessiva** se esiste un oggetto $\Omega \in \mathrm{Ob}(\mathscr{C})$ e un funtore:

$$\Sigma: \mathscr{C} \to \mathscr{C}$$

— cioè un endofuntore di $\mathscr{C}$ —, insieme a una trasformazione naturale:

$$\delta: \mathrm{Id}_{\mathscr{C}} \Rightarrow \Sigma$$

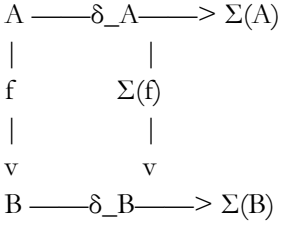
tale che per ogni oggetto A il morfismo $\delta_A: A \to \Sigma(A)$ sia un **morfismo di auto-presentazione**: $\Sigma(A)$ rappresenta A come oggetto della categoria, ossia c'è un isomorfismo canonico:

$$\Sigma(A) \cong \mathrm{Hom}_{\mathscr{C}}(A, \Omega)$$

nella categoria degli insiemi, dove Ω gioca il ruolo di **oggetto classificante** — generalizzazione categoriale del classificatore di sottooggetti dei topos.

Una categoria con questa struttura è tale che ogni suo oggetto si *presenta a se stesso* attraverso il funtore Σ: ogni stato ha, nella struttura della categoria, una rappresentazione di se stesso come stato. Questo è un modello astratto dell'autoriflessività fenomenica — del fatto che gli stati coscienti non soltanto hanno un contenuto intenzionale diretto verso il mondo, ma si presentano a se stessi nel loro presentare il mondo: la coscienza è, nella terminologia husserliana, sempre insieme *intenzionalità* e *auto-donazione* (Selbstgegebenheit).

La condizione di naturalità di δ — la commutatività dei quadrati:



per ogni morfismo f: A → B — esprime il fatto che l'auto-presentazione è **coerente** con la struttura morfologica: se A si trasforma in B attraverso f, allora l'auto-presentazione di A si trasforma coerentemente nell'auto-presentazione di B attraverso Σ(f). Non vi è salto o discontinuità nell'auto-presentazione lungo i morfismi: ogni transizione cognitiva trasporta la propria auto-donazione in modo strutturalmente coerente.

Definiamo ora: una categoria $\mathscr{C}$ è **fenomenicamente chiusa** se è intrinsecamente autoriflessiva e se, in aggiunta, il funtore di auto-presentazione Σ è **conservativo**: ogni morfismo f: A → B tale che Σ(f): Σ(A) → Σ(B) sia un isomorfismo è esso stesso un isomorfismo. In altri termini: se l'auto-presentazione di due stati è indistinguibile, i due stati sono identici.

La chiusura fenomenica cattura, formalmente, la tesi che il *come ci si presenta* determina il *che cosa si è*, almeno per gli stati coscienti: due esperienze che si presentino identicamente a se stesse sono la medesima esperienza. Questo non è un principio banale: esso esclude la possibilità degli **zombi fenomenici alla rovescia** — sistemi che abbiano strutture di auto-presentazione identiche ma esperienze diverse. La chiusura fenomenica è la tesi che tale possibilità è categorialmente incoerente: nella categoria degli stati coscienti, la struttura di auto-presentazione è un invariante completo.

Si può ora formulare il problema difficile nella sua versione categoriale più precisa. Sia $\mathfrak{M}$ la categoria degli stati mentali con la propria struttura inferenziale, e sia $\mathscr{P}$hys la categoria degli stati fisici. Il problema difficile è la congettura che $\mathfrak{M}$ **sia fenomenicamente chiusa ma $\mathscr{P}$hys no**. In altri termini: nella categoria degli stati fisici non esiste, in generale, un endofuntore di auto-presentazione che soddisfi le condizioni di conservatività richieste. I sistemi fisici non si «presentano a se stessi» nel senso pertinente; essi semplicemente si trovano in stati che producono effetti causali su altri stati.

La chiusura fenomenica di $\mathfrak{M}$, d'altro canto, è esattamente ciò che caratterizza la coscienza come coscienza: il fatto che i suoi stati abbiano la struttura formale di un oggetto classificante interno, che ogni stato sia la propria rappresentazione nel senso tecnicamente preciso definito sopra. Un sistema fisico che non abbia questa struttura sarà per quanto si voglia funzionalmente equivalente a un sistema cosciente — potrà soddisfare la condizione di sopravvenienza, avere la stessa struttura causa-effetto misurata dalla IIT — e nondimeno non possiederà la chiusura fenomenica.

XIII. Il topos degli stati intenzionali e la logica interna dell'esperienza

La nozione di chiusura fenomenica raggiunge il suo sviluppo naturale nel contesto dei **topos**, una delle strutture più ricche e sorprendenti della matematica contemporanea. Un topos elementare — nella definizione di Lawvere e Tierney — è una categoria $\mathscr{T}$ che soddisfa tre condizioni: possiede tutti i limiti finiti, possiede un classificatore di sottooggetti Ω (un oggetto tale che i sottooggetti di ogni oggetto A corrispondano in

modo naturale ai morfismi $A \to \Omega$), e possiede esponenziali (oggetti B^A che rappresentano i morfismi da A a B).

La caratteristica fondamentale dei topos è che essi hanno una **logica interna**: ad ogni topos è associata una logica intuizionista — in generale non classica — che descrive le relazioni tra i sottooggetti degli oggetti del topos. Il classificatore Ω è l'«oggetto di verità», e i suoi elementi sono i valori di verità della logica interna: in **Set** (che è un topos), $\Omega = \{0, 1\}$ e la logica è quella classica bivalente; in un topos di fasci su uno spazio topologico, Ω è l'insieme degli aperti, e la logica interna è la logica dei fasci — intuizionista, con la proprietà che la verità di un'affermazione è sempre locale (dipende da quale aperto si considera).

Se $\mathfrak{M}$ è un topos, la sua logica interna descrive le relazioni inferenziali tra stati mentali dal punto di vista interno al sistema cognitivo: non la logica classica osservata dall'esterno, ma la logica vissuta dall'interno. Ed è significativo che questa logica sia generalmente intuizionista: la legge del terzo escluso — che per ogni proposizione A valga $A \vee \neg A$ — non è in generale valida nella logica interna di un topos. Questo risuona con la fenomenologia dell'incertezza cognitiva: nella vita mentale concreta, molte questioni si trovano in uno stato di indeterminazione genuina, non risolvibile a piacere per via di un principio logico esterno.

L'oggetto classificante Ω del topos $\mathfrak{M}$ è, nella lettura proposta, l'oggetto di auto-presentazione che definisce la chiusura fenomenica: esso è la struttura che porta ogni stato a «classificarsi» come stato della categoria. I sottooggetti di ogni stato mentale M — i suoi «aspetti» o «componenti» fenomenicamente rilevanti — corrispondono ai morfismi $M \to \Omega$, e la totalità di questi morfismi costituisce la struttura fenomenica di M.

La **metrica dell'esperienza** — quella struttura che nella sezione VII si era indicata come una varietà fenomenica $\mathscr{P}$ da costruire — può ora essere intesa come la struttura geometrica sull'oggetto Ω del topos $\mathfrak{M}$. Se Ω è dotato di una topologia interna (che ha senso nei topos di fasci), la distanza tra esperienze è la distanza interna sullo spazio degli aperti di Ω — una distanza che generalizza, in modo intrinseco e non esterno, la metrica di Fisher-Rao della geometria dell'informazione. La relazione tra le due strutture — quella statistica e quella interna al topos — è un problema aperto di considerevole difficoltà e interesse.

XIV. Il limite come identità: Yoneda rivisitato attraverso l'esperienza

Si giunga ora alla tesi conclusiva, che riporta al lemma di Yoneda ma lo reinterpreta alla luce della struttura di chiusura fenomenica.

Il lemma di Yoneda afferma che un oggetto A è determinato dal funtore $\mathrm{Hom}(A, -)$. Nella sezione VI si era osservato che questo non esclude la possibilità di differenze extra-morfologiche tra oggetti. Ma se $\mathfrak{M}$ è fenomenicamente chiusa — se il funtore di auto-presentazione Σ è conservativo —, allora ogni differenza rilevante si manifesta

necessariamente nella struttura di auto-presentazione. Combinando le due condizioni, si ottiene:

Un oggetto A in una categoria fenomenicamente chiusa $\mathfrak{M}$ è determinato non soltanto dal funtore Hom(A, -) — la sua posizione nella rete morfologica — ma anche dal morfismo $\delta_A: A \to \Sigma(A)$ — la sua auto-presentazione. Due oggetti sono identici se e soltanto se sono isomorfi come oggetti morfologici (Yoneda) e le loro auto-presentazioni sono compatibili con tale isomorfismo.

Questa è la struttura formale dell'identità fenomenica: non la semplice identità strutturale (Leibniz), non la semplice identità qualitativa (qualia irriducibili), ma l'identità che emerge dalla coincidenza di struttura morfologica e struttura di auto-presentazione — ciò che si potrebbe chiamare l'**identità come limite**, nel senso che l'oggetto è il limite universale del suo stesso funtore di auto-presentazione.

Il problema difficile della coscienza, in questa prospettiva, è la domanda se $\mathfrak{M}$ sia effettivamente fenomenicamente chiusa — se esista davvero un endofuntore Σ di auto-presentazione che soddisfi le condizioni richieste. È una domanda che non si può rispondere dalla sola analisi concettuale: richiede un'indagine empirica sulla struttura degli stati coscienti che vada molto più in profondità di quanto la neuroscienza attuale permetta. Ma la sua formulazione precisa è già un progresso rispetto all'impasse tra riduzionismo e dualismo: essa mostra che il fenomenico non è una sostanza separata (come nel dualismo cartesiano) né un epifenomeno irrilevante (come nel riduzionismo eliminativista), ma una struttura categoriale specifica — la chiusura dell'auto-presentazione — che alcuni sistemi fisici potrebbero realizzare e altri no.

XV. Conclusioni: il formalismo come cartografia dell'incompiuto

Il percorso compiuto attraverso la teoria delle categorie, la geometria dell'informazione e la logica dei topos non ha prodotto una soluzione al problema difficile della coscienza. Esso ha prodotto qualcosa di diverso e, a giudizio di chi scrive, di più onesto: una **cartografia precisa dell'incompiuto** — una mappa che indica con esattezza dove si trovano i confini del sapere formale, e che forma di struttura avrebbe la risposta cercata se mai si riuscisse a trovarla.

I risultati positivi dell'indagine sono tre. Il primo è la formalizzazione della distinzione fraseiana senso/riferimento come distinzione tra la classe di equivalenza di un funtore e il suo valore su oggetti specifici: un guadagno di precisione semantica che le teorie proposizionali classiche non erano in grado di ottenere. Il secondo è l'identificazione della struttura di aggiunzione come modello del ciclo percezione-azione: un risultato che connette la teoria delle categorie con le neuroscienze computazionali in modo non metaforico ma formalmente preciso. Il terzo è la nozione di categoria fenomenicamente chiusa come condizione formale dell'auto-presentazione: un costrutto che reincornicia il problema difficile in termini che, a differenza delle formulazioni tradizionali, consentono almeno di stabilire che tipo di fatto empirico potrebbe risolverlo.

Il risultato negativo è uno e fondamentale: il lemma di Yoneda, pur nella sua profondità, non esclude la possibilità di differenze extra-morfologiche tra oggetti, e quindi non può, da solo, escludere i qualia come residuo irriducibile. L'olismo morfologico che il lemma di Yoneda impone è una condizione necessaria per l'identità fenomenica — due esperienze morfologicamente indistinguibili sono certamente identiche — ma non sufficiente: rimane aperta la possibilità che esperienze morfologicamente equivalenti siano fenomenicamente distinte, e questa è precisamente la forma tecnica della tesi degli invertiti qualitativi.

Ciò che emerge dall'intera analisi è una prospettiva che si potrebbe denominare **strutturalismo fenomenologico selettivo**: la struttura categoriale cattura tutto ciò che è comunicabile, condivisibile, linguisticamente articolabile dell'esperienza — tutto ciò che può diventare oggetto di inferenza, di accordo intersoggettivo, di revisione razionale. Ciò che essa non cattura — se davvero non lo cattura, e questa rimane una questione aperta — è esattamente ciò che è non-comunicabile nell'esperienza: non perché misterioso in linea di principio, ma perché privo di morfismi verso qualsiasi altro stato che non ne condivida la struttura di auto-presentazione.

Il silenzio del formalismo al cospetto dei qualia non è un difetto del formalismo: è la sua risposta più onesta.

Categorie arricchite, intenzionalità graduata e la metrica del contenuto

---

XVI. Il limite dell'approccio Set-arricchito e la necessità dell'arricchimento

Tutta l'analisi svolta fino a questo punto ha tacitamente presupposto che gli spazi di morfismi $\mathrm{Hom}_{\mathfrak{M}}(M, N)$ tra stati mentali fossero insiemi — oggetti della categoria **Set**. Questa assunzione, che è standard nella teoria delle categorie elementare, nasconde una semplificazione che diventa problematica non appena si prenda sul serio la struttura quantitativa dell'esperienza cognitiva. Le transizioni tra stati mentali non sono semplicemente presenti o assenti — non si tratta di morfismi che ci sono o non ci sono —: esse hanno una **forza**, una **probabilità**, una **facilità relativa** che è parte essenziale della struttura cognitiva. Credere fermamente che p e credere debolmente che p non sono lo stesso stato mentale, e la transizione inferenziale da «forse q» a «certamente q» non è morfologicamente equivalente alla transizione da «certamente ¬q» a «certamente q». La teoria delle categorie ordinarie, con i suoi spazi di morfismi discreti, non ha gli strumenti per catturare questa gradazione.

La risposta tecnica a questa limitazione si chiama **categoria arricchita** (*enriched category*), e fu sviluppata sistematicamente da Gregory Maxwell Kelly nel corso degli anni Settanta, con antecedenti importanti nel lavoro di Eilenberg e Kelly del 1966. L'idea di base è di sostituire **Set** — la categoria in cui vivono gli spazi di morfismi ordinari — con una categoria monoïdale chiusa $\mathscr{V}$ più strutturata, i cui oggetti

possono essere spazi metrici, spazi vettoriali, reticoli, spazi di probabilità o qualsiasi altra struttura dotata di un'operazione di «tensore» che permetta di comporre i morfismi.

Una **categoria monoidale** $(\mathcal{V}, \otimes, I)$ è una categoria dotata di un funtore bifuntoriale $\otimes: \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ (il prodotto tensoriale), di un oggetto unità $I \in \mathrm{Ob}(\mathcal{V})$, e di isomorfismi naturali di associatività e unità che soddisfano le equazioni di coerenza di Mac Lane:

$\alpha_{A,B,C}: (A \otimes B) \otimes C \cong A \otimes (B \otimes C)$ (associatore)
$\lambda_A: I \otimes A \cong A$ (unità sinistra)
$\varrho_A: A \otimes I \cong A$ (unità destra)

soggetti al **pentagonale** e al **triangolare**:

$\alpha_{A,B,C\otimes D} \circ \alpha_{A\otimes B,C,D} = (\mathrm{id}_A \otimes \alpha_{B,C,D}) \circ \alpha_{A,B\otimes C,D} \circ (\alpha_{A,B,C} \otimes \mathrm{id}_D)$
$(\mathrm{id}_A \otimes \lambda_B) \circ \alpha_{A,I,B} = \varrho_A \otimes \mathrm{id}_B$

Una categoria monoïdale è **chiusa** se per ogni oggetto B il funtore $- \otimes B: \mathcal{V} \to \mathcal{V}$ ha un aggiunto destro $[B, -]: \mathcal{V} \to \mathcal{V}$, ossia se esiste una biiezione naturale:

$\mathrm{Hom}_{\mathcal{V}}(A \otimes B, C) \cong \mathrm{Hom}_{\mathcal{V}}(A, [B, C])$

L'oggetto $[B, C]$ è l'**oggetto interno di morfismi** da B a C, e la sua esistenza garantisce che $\mathcal{V}$ abbia una nozione interna di «spazio delle mappe» tra oggetti, non soltanto una nozione esterna (data dalla classe $\mathrm{Hom}_{\mathcal{V}}(B, C)$).

Una **$\mathcal{V}$-categoria** $\mathcal{C}$ arricchita su $\mathcal{V}$ è allora una struttura con una classe di oggetti $\mathrm{Ob}(\mathcal{C})$ e, per ogni coppia (A, B) di oggetti, un **oggetto di morfismi** $\mathcal{C}(A, B) \in \mathrm{Ob}(\mathcal{V})$ — non più un insieme, ma un oggetto della categoria di arricchimento —, insieme a morfismi di composizione e di identità in $\mathcal{V}$:

$\circ_{A,B,C}: \mathcal{C}(B, C) \otimes \mathcal{C}(A, B) \to \mathcal{C}(A, C)$
$j_A: I \to \mathcal{C}(A, A)$

soggetti alle ovvie condizioni di associatività e unitarietà espresse come commutazione di diagrammi in $\mathcal{V}$.

XVII. La categoria degli stati mentali arricchita su $[0,\infty]$

L'esempio di arricchimento più rilevante per la teoria dell'intenzionalità è il caso $\mathcal{V} =$ **$([0,\infty], \geq, +, 0)$**: la categoria monoïdale con oggetti che sono i numeri reali non

negativi (con ∞), i cui morfismi da a a b esistono se e soltanto se a ≥ b (ossia il solo morfismo non banale va dal numero più grande al più piccolo), il cui prodotto tensoriale è la somma ordinaria + e il cui oggetto unità è 0. Una categoria arricchita su [0,∞] è esattamente uno **spazio metrico generalizzato** nel senso di Lawvere (1973): gli oggetti sono i punti, e $\mathscr{C}(A, B) \in [0,\infty]$ è la distanza da A a B, con la condizione di composizione:

$$\mathscr{C}(B, C) + \mathscr{C}(A, B) \geq \mathscr{C}(A, C)$$

che è la disuguaglianza triangolare, e la condizione di identità:

$$0 = \mathscr{C}(A, A)$$

La struttura non richiede la simmetria — $\mathscr{C}(A, B) = \mathscr{C}(B, A)$ — né la separazione — $\mathscr{C}(A, B) = 0 = \mathscr{C}(B, A)$ implica A = B. Essa è dunque più generale della nozione standard di spazio metrico, e include gli spazi quasimetrici, i sistemi di costo di trasporto ottimale, e le strutture geometriche delle categorie di stati di un automa probabilistico.

Considero ora la categoria degli stati mentali $\mathfrak{M}$ arricchita su [0,∞]. Il valore $\mathfrak{M}(M, N) \in [0,\infty]$ è la **distanza cognitiva** da M a N: il costo minimo della transizione inferenziale o associativa che porta dallo stato M allo stato N. Questa distanza non è in generale simmetrica: il costo di inferire N a partire da M può essere molto diverso dal costo dell'inferenza inversa. Il fatto che «è piovuto ieri» implichi immediatamente «c'era umidità ieri» (distanza cognitiva piccola) non significa che «c'era umidità ieri» implichi con la stessa facilità «è piovuto ieri» (distanza cognitiva grande, perché esistono molte altre spiegazioni dell'umidità).

In questa struttura arricchita, il **lemma di Yoneda** si generalizza in modo netto. Il funtore di Yoneda arricchito:

$$y: \mathfrak{M} \to \mathscr{V}\text{-Fun}(\mathfrak{M}^{\{op\}}, \mathscr{V})$$

porta ogni oggetto M al $\mathscr{V}$-funtore $\mathfrak{M}(-, M): \mathfrak{M}^{\{op\}} \to [0,\infty]$ che assegna a ogni N la distanza $\mathfrak{M}(N, M)$. La versione arricchita del lemma afferma che per ogni $\mathscr{V}$-funtore $F: \mathfrak{M} \to [0,\infty]$ esiste un isomorfismo in $\mathscr{V}$:

$$\mathscr{V}\text{-Nat}(\mathfrak{M}(M, -), F) \cong F(M)$$

dove ora $\mathscr{V}$-Nat indica le trasformazioni naturali arricchite — quelle per le quali la condizione di naturalità è espressa attraverso morfismi in $\mathscr{V}$ piuttosto che eguaglianze tra funzioni d'insiemi.

La traduzione in termini cognitivi è illuminante. $F(M) \in [0,\infty]$ è il «costo» o la «distanza» associata allo stato M da una certa valutazione F — si pensi a F come alla

funzione di credibilità, o di attivazione, o di sorpresa (surprisal) che assegna a ogni stato il proprio valore informazionale. Il lemma di Yoneda arricchito afferma che questo valore è completamente determinato dalla struttura delle distanze cognitive tra M e tutti gli altri stati: non serve informazione aggiuntiva su M oltre alla sua posizione nella geometria cognitiva di $\mathfrak{M}$.

Questa è una versione metricamente precisa del principio brandomiano del contenuto inferenziale: il contenuto di uno stato mentale è la sua posizione nello spazio metrico delle inferenze cognitive, e nessun'altra caratterizzazione è necessaria — a meno che, come si è argomentato, non vi sia un residuo fenomenico extra-morfologico.

XVIII. La connessione di Levi-Civita nella varietà dei contenuti intenzionali

L'arricchimento su $[0,\infty]$ trasforma la categoria degli stati mentali in uno spazio metrico generalizzato. Ma uno spazio metrico non ha ancora la struttura sufficiente per definire la nozione di **trasporto parallelo** — la capacità di spostare vettori tangenti da un punto all'altro lungo una curva in modo «senza rotazione». Questa struttura, che nella geometria riemanniana è catturata dalla connessione di Levi-Civita, è necessaria per rispondere a domande del tipo: quando un sistema cognitivo sposta la propria attenzione da un contenuto M a un contenuto N vicino, come si trasformano i giudizi che esso formulava riguardo ai correlati di M in giudizi riguardo ai correlati di N?

Supponiamo che la varietà statistica dei contenuti intenzionali — parametrizzata da $\theta \in \Theta \subseteq \mathbb{R}^n$ come nella sezione VII — abbia la metrica di Fisher-Rao g_{ij}(θ). La connessione di Levi-Civita associata a questa metrica è la connessione affine ∇ caratterizzata da due proprietà: la compatibilità con la metrica e l'assenza di torsione. Queste condizioni determinano i simboli di Christoffel:

$$\Gamma^{k}_{ij}(\theta) = (1/2)\, g^{kl}(\theta)\, (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij})$$

dove g^{kl} è l'inversa del tensore metrico, e gli indici ripetuti seguono la convenzione di Einstein per la somma. Il **trasporto parallelo** di un vettore $V = V^i(\partial/\partial\theta^i)$ lungo una curva $\gamma: [0,1] \to \Theta$ con $\gamma(0) = \theta_0$, $\gamma(1) = \theta_1$ è determinato dall'equazione differenziale:

$$dV^k/dt + \Gamma^{k}_{ij}(\gamma(t)) \cdot (d\gamma^i/dt) \cdot V^j(t) = 0$$

con condizione iniziale $V(0) = V_0$. La soluzione $V(1)$ è il vettore trasportato parallelamente da θ_0 a θ_1 lungo γ: esso conserva la norma (per la compatibilità con la metrica) ma ruota secondo la curvatura dello spazio.

In termini cognitivi, il trasporto parallelo descrive come un sistema di credenze o di aspettative (un vettore tangente nello spazio dei contenuti) si modifica quando il soggetto «sposta il contesto» — passa da un insieme di presupposti a un altro vicino.

La curvatura dello spazio dei contenuti intenzionali — il tensore di Riemann R^k_{lij}(θ) — misura quanto questo trasporto dipenda dal cammino percorso: se la curvatura è nulla (spazio piatto), il risultato del trasporto è indipendente dal percorso, e la struttura cognitiva è «integrabile» nel senso che si può definire globalmente un sistema di coordinate coerente. Se la curvatura è non nulla, trasportare le stesse credenze lungo percorsi diversi dà risultati diversi: l'ordine in cui si affrontano i contesti cognitivi conta, e la struttura delle credenze non è ricostruibile da un sistema globale di riferimento.

Questa non-integrabilità ha un correlato fenomenologico preciso: è ciò che Husserl chiama l'**orizzonte** del vissuto — il fatto che ogni esperienza si dà sempre in un contesto di rimandi che cambia quando si cambia prospettiva, e che non è possibile ricostituire l'intero orizzonte da nessun punto di vista privilegiato. La curvatura della varietà dei contenuti intenzionali è la misura geometrica della «prospettivicità» dell'esperienza: della dipendenza dal contesto che è costitutiva della vita cognitiva concreta.

Esplicitiamo la formula per la curvatura scalare — la contrazione completa del tensore di Riemann — che riassume in un singolo numero la «curvatura media» dello spazio dei contenuti:

$$R(\theta) = g^{ij}(\theta)\, R_{i\square}(\theta) = g^{ij}(\theta)\, R^{k}{}_{i\square\square}(\theta)$$

dove $R_{i\square} = R^{k}{}_{i\square\square}$ è il tensore di Ricci. Per una famiglia esponenziale bidimensionale — che è il modello più semplice di uno spazio di contenuti intenzionali con due gradi di libertà — si può calcolare esplicitamente che $R = 0$: le famiglie esponenziali danno luogo a varietà statistiche piatte. Questo significa che il cervello, nella misura in cui i suoi stati generativi si modellano come distribuzioni di una famiglia esponenziale (gaussiane, binomiali, Poisson), opera in uno spazio cognitivo localmente piatto: il trasporto parallelo è integrabile e la struttura delle credenze è consistente indipendentemente dall'ordine di acquisizione delle evidenze. Ciò è coerente con l'ipotesi del cervello predittivo bayesiano: la coerenza dell'inferenza bayesiana richiede precisamente la piattezza dello spazio dei parametri rispetto alla connessione duale di Amari — la connessione ± 1 della geometria dell'informazione, che è distinta dalla connessione di Levi-Civita ma coincide con essa nel caso piatto.

Tuttavia, non tutti i contenuti intenzionali si modellano come famiglie esponenziali. Le credenze su strutture relazionali complesse, le aspettative su sequenze temporali, i sistemi di preferenze con non-linearità — tutti questi casi generano varietà statistiche con curvatura non nulla, e quindi strutture cognitive non-integrabili. In questi casi, la logica dell'esperienza non è quella dell'aggiornamento bayesiano standard: è una logica dipendente dal percorso, che la sola struttura probabilistica non basta a descrivere.

XIX. Enriquecimiento sobre categorías de probabilidad: la $\mathscr{M}$-categoría de estados mentales

Il passo successivo nell'arricchimento della struttura è di considerare come categoria di arricchimento non il semianello $([0,\infty], +, 0)$ ma una categoria di spazi di probabilità — in modo da catturare non solo le distanze tra stati ma le intere distribuzioni di probabilità sulle transizioni. La struttura naturale che emerge è quella delle **categorie di Markov**, introdotte da Cho e Jacobs (2019) e sviluppate nel quadro della **probabilità categoriale** da Fritz (2020) attraverso la nozione di categoria di sintesi (synthetic probability theory).

Una **categoria di Markov** è una categoria simmetrica monoïdale $\mathscr{C}$ nella quale ogni oggetto A è dotato di un comonoide commutativo — morfismi di copia $\delta_A: A \to A \otimes A$ e di cancellazione $\varepsilon_A: A \to I$ — che si comportano coerentemente con la struttura monoïdale. I morfismi di una categoria di Markov si interpretano come **kernel markoviani**: mappe stocastiche che portano distribuzioni su A in distribuzioni su B, generalizzando le funzioni deterministiche al caso probabilistico.

L'esempio fondamentale è la categoria **FinStoch** i cui oggetti sono gli insiemi finiti e i cui morfismi da X a Y sono le matrici stocastiche $M: X \times Y \to [0,1]$ con $\Sigma_{y \in Y} M(x, y) = 1$ per ogni $x \in X$. La composizione di due kernel $M: X \to Y$ e $N: Y \to Z$ è la matrice prodotto stocastico:

$$(N \circ M)(x, z) = \Sigma_{y \in Y} N(y, z) \cdot M(x, y)$$

che è la legge di Chapman-Kolmogorov: la probabilità di arrivare in z partendo da x, passando per y con probabilità M(x,y) e poi andando a z con probabilità N(y,z).

Arricchire la categoria degli stati mentali $\mathfrak{M}$ su FinStoch significa assegnare a ogni coppia (M, N) di stati mentali non un singolo numero — la distanza cognitiva — ma un **kernel stocastico** $\mathfrak{M}(M, N) \in \mathrm{Hom}_{\mathrm{FinStoch}}(S_M, S_N)$, dove S_M e S_N sono gli insiemi di «configurazioni sensoriali» compatibili con M e N rispettivamente. Il valore $\mathfrak{M}(M, N)(s_M, s_N)$ è la probabilità che, trovandosi nello stato M e osservando la configurazione sensoriale s_M, si transizioni verso lo stato N con la configurazione s_N.

Questa struttura arricchita cattura in modo naturale l'**incertezza cognitiva** come caratteristica strutturale della categoria: non soltanto gli stati mentali hanno contenuti probabilistici (come nella teoria bayesiana), ma le transizioni tra stati sono esse stesse di natura stocastica, con distribuzioni sulle possibili configurazioni di arrivo. Il lemma di Yoneda in questa categoria arricchita — la cui formulazione precisa richiede la teoria dei prefasci su categorie di Markov, ancora in sviluppo nella letteratura — affermerebbe che ogni stato mentale è determinato dalla sua distribuzione stocastica di morfismi verso tutti gli altri stati: non dalla sua posizione nello spazio metrico degli stati, ma dalla struttura delle sue *transizioni probabilistiche* nell'intera rete cognitiva.

Questa è, in termini matematicamente precisi, la tesi che la psicologia computazionale bayesiana sostiene in termini informali: il contenuto cognitivo di uno stato è la distribuzione di probabilità sulle esperienze future che esso predice, congiuntamente

alla distribuzione sulle esperienze passate che lo hanno generato. La struttura categoriale di Markov è la struttura astratta che questa tesi presuppone.

La struttura simpliciale del tempo, il nervo di una categoria e il flusso dell'esperienza

---

XX. Il problema del tempo fenomenico e le sue radici husserliane

C'è un aspetto dell'esperienza cosciente che le sezioni precedenti hanno deliberatamente lasciato in ombra, e che non è possibile eludere più a lungo senza che l'analisi resti irrimediabilmente parziale: la **struttura temporale** del vissuto. Ogni esperienza non si dà come un punto istantaneo — una configurazione statica di stati mentali simultanei: essa si estende nel tempo, porta con sé la traccia di ciò che l'ha preceduta e l'anticipazione di ciò che verrà, e questa struttura di protensione e ritenzione — per usare la terminologia husserliana delle *Lezioni sulla fenomenologia della coscienza interna del tempo* — non è un ornamento accessorio dell'esperienza ma la sua forma costitutiva.

Husserl distingue tre livelli temporali nell'esperienza: il **pronto presente** (Urimpression), che è la fase di originaria presentazione del contenuto attuale; la **ritenzione**, che è la trattenuta modificata dei momenti appena trascorsi, ancora vivente nel presente come eco strutturale; e la **protensione**, che è l'anticipazione vuota dei momenti immediatamente futuri, che il presente porta con sé come orizzonte aperto. La struttura temporale non è dunque una successione discreta di istanti che si escludono a vicenda, ma una continuità di compresenza stratificata: il presente ritiene il passato prossimo e protende verso il futuro prossimo in un unico atto di coscienza.

La domanda che si pone è: questa struttura ha un correlato formale nella teoria delle categorie? È possibile costruire una struttura categoriale che catturi non soltanto le relazioni sincroniche tra stati mentali — le transizioni inferenziali di $\mathfrak{M}$ — ma anche la loro articolazione diacronica: il modo in cui si dispongono, si trasformano e si conservano nel tempo?

La risposta tecnica coinvolge uno dei ponti più eleganti della matematica contemporanea: la teoria **simpliciale** e, in particolare, la costruzione del **nervo** di una categoria.

XXI. Insiemi simpliciali e la topologia del tempo cognitivo

Un **simplesso standard** di dimensione n, indicato Δ^n, è la realizzazione geometrica convessa di n+1 punti in posizione generale in $\mathbb{R}^n$:

$$\Delta^n = \{ (t_0, t_1, ..., t) \in \mathbb{R}^{n+1} : \Sigma_i \, t_i = 1, t_i \geq 0 \}$$

Il simplesso Δ^0 è un punto, Δ^1 è un segmento, Δ^2 è un triangolo, Δ^3 è un tetraedro. Le **mappe faccia** δ^i: $\Delta^{n-1} \to \Delta^n$ per i = 0, ..., n includono Δ^{n-1} come la i-esima faccia di Δ^n omettendo la coordinata i-esima; le **mappe degenerazione** σ^i: $\Delta^{n+1} \to \Delta^n$ proiettano Δ^{n+1} su Δ^n collassando le coordinate i e i+1.

Un **insieme simpliciale** è un funtore controvariante X: Δ^{op} → Set dalla **categoria simpliciale** Δ — la categoria degli ordinali finiti non vuoti [0], [1], [2], ... con morfismi le mappe debolmente crescenti — verso la categoria degli insiemi. Concretamente, X è una sequenza di insiemi {X }_{n≥0} con mappe faccia:

d_i: X $\to$ X $_{-1}$, i = 0, 1, ..., n

e mappe degenerazione:

s_i: X $\to$ X $_{+1}$, i = 0, 1, ..., n

soggetti alle **identità simpliciali**:

$d_i \circ d = d_{-1} \circ d_i$ per i < j
$s_i \circ s = s_{+1} \circ s_i$ per i ≤ j
$d_i \circ s = s_{-1} \circ d_i$ per i < j
$d_i \circ s = id$ per i = j oppure i = j+1
$d_i \circ s = s \circ d_{i-1}$ per i > j+1

XXII. Il nervo di una categoria e la sua struttura cognitiva

Il collegamento tra insiemi simpliciali e teoria delle categorie è fornito dalla costruzione del **nervo**. Dato una piccola categoria $\mathscr{C}$, il suo nervo N$\mathscr{C}$ è l'insieme simpliciale:

N$\mathscr{C}$ = { catene composibili di n morfismi in $\mathscr{C}$: $A_0 \to^{f_1} A_1 \to^{f_2} \cdots \to^{f} A$ }

L'insieme degli n-simplessi è dunque l'insieme delle sequenze di n morfismi consecutivi. Le mappe faccia agiscono come:

$d_0(A_0 \to^{f_1} A_1 \to \cdots \to A) = (A_1 \to^{f_2} \cdots \to A)$
$d (A_0 \to \cdots \to A_{-1} \to^{f} A) = (A_0 \to \cdots \to A_{-1})$
$d_i(A_0 \to \cdots \to A) = (A_0 \to \cdots \to A_{i-1} \to^{f_{i+1} \circ f_i} A_{i+1} \to \cdots \to A)$ per 0 < i < n

La i-esima mappa faccia intermedia compone dunque i morfismi f_i e f_{i+1} in un unico morfismo $f_{i+1} \circ f_i$. Le mappe degenerazione inseriscono morfismi identità:

$$s_i(A_0 \to \cdots \to A) = (A_0 \to \cdots \to A_i \to^{id_{A_i}} A_i \to \cdots \to A)$$

Un risultato fondamentale di Grothendieck afferma che un insieme simpliciale X è isomorfo al nervo di una categoria se e soltanto se soddisfa la **condizione di Segal**: per ogni $n \geq 2$, la mappa di restrizione alle spigoli:

$$X \to X_1 \times_{X_0} X_1 \times_{X_0} \cdots \times_{X_0} X_1 \quad \text{(n fattori)}$$

è una biiezione. La condizione di Segal afferma che ogni catena di morfismi componibili è determinata in modo unico dai suoi passi elementari — è la condizione di unicità della composizione.

XXIII. La struttura simpliciale dell'esperienza temporale

Un n-simplesso di $N\mathfrak{M}$ — una catena di n transizioni cognitive consecutive:

$$M_0 \to^{f_1} M_1 \to^{f_2} M_2 \to \cdots \to^{f} M$$

— si interpreta come un **episodio cognitivo di profondità n**: una sequenza di n passi di elaborazione mentale che porta dallo stato M_0 allo stato M attraverso gli stati intermedi.

La mappa faccia d_0 «dimentica» l'inizio dell'episodio e lascia la ritenzione: la catena $M_1 \to \cdots \to M$ ancora presente nella coscienza come traccia strutturale. La mappa d dimentica l'orizzonte protenzionale e lascia la struttura già vissuta $M_0 \to \cdots \to M_{-1}$. La mappa faccia intermedia d_i compone due transizioni contigue in un'unica transizione percepita come unitaria — il processo di **sintesi cognitiva** che integra due passi elementari in un unico atto di coscienza.

Questa lettura formalizza in termini precisi la struttura husserliana della coscienza temporale. La ritenzione non è il ricordo — che sarebbe un atto separato di rimemorazione —, ma la presenza modificata del passato nel presente, esattamente come d_0 non cancella l'informazione sull'inizio dell'episodio ma la riposiziona: M_1 è ancora strutturalmente presente nella catena risultante, ma non più nella posizione di punto di partenza assoluto. La protensione corrisponde alla degenerazione s che inserisce un morfismo identità al termine della catena, aprendo lo spazio per la continuazione senza ancora determinarla.

La condizione di Segal è in questo contesto la condizione di **composizionalità cognitiva**: ogni episodio di n passi è determinato in modo unico dai suoi passi

elementari e dalla struttura delle loro composizioni. Essa è la versione simpliciale del principio fregeano di composizionalità semantica.

XXIV. ∞-categorie, omotopia cognitiva e la coerenza dell'esperienza

Il passo più arduo è di abbandonare la richiesta che la condizione di Segal sia soddisfatta con biiezioni esatte, sostituendola con la condizione più debole di **equivalenze omotopiche**. Questa generalizzazione porta dalla teoria delle categorie alla teoria delle **∞-categorie** — o quasi-categorie — associata ai lavori di André Joyal e Jacob Lurie.

Una **quasi-categoria** (nel senso di Joyal) è un insieme simpliciale X che soddisfa la condizione del **corno interno**: per ogni $0 < k < n$ e ogni mappa di corno interno $\Lambda^n \to X$ (dove $\Lambda^n = \Delta^n$ privato dell'interno e della faccia k-esima), esiste almeno un prolungamento $\Delta^n \to X$. Non è richiesta l'unicità. In termini cognitivi: ogni catena di $n-1$ morfismi composibili con una lacuna in posizione k può essere completata in almeno un modo, e modi diversi di completarla sono considerati equivalenti se vi è un $(n+1)$-simplesso che li connette — ossia se sono **cognitivamente omotopi**.

La non-unicità della composizione è il punto strutturalmente decisivo. Nelle categorie ordinarie la composizione è esatta; nelle ∞-categorie essa è definita soltanto a meno di omotopia — una nozione di equivalenza più debole dell'isomorfismo, che permette di comporre morfismi in modi diversi purché questi modi siano connessi da una struttura gerarchica infinita di coerenze. I 2-morfismi (omotopie tra composizioni), i 3-morfismi (omotopie tra omotopie), e così via ad ogni ordine, formano una struttura che codifica tutti i gradi della coerenza composizionale.

L'interpretazione cognitiva è la seguente. Nella vita mentale concreta, la composizione di due atti cognitivi non è mai esattamente determinata in modo unico: vi sono sempre percorsi inferenziali leggermente diversi che portano dallo stesso stato di partenza allo stesso stato di arrivo, e questi percorsi sono fra loro **cognitivamente equivalenti** nella misura in cui producono lo stesso stato finale ai fini delle inferenze e delle azioni successive. La struttura delle ∞-categorie cattura precisamente questa **tolleranza cognitiva**: la coerenza dell'esperienza non è rigidamente combinatoria ma omotopicamente flessibile.

Questa flessibilità ha un correlato neuroscientificamente plausibile. I circuiti neurali sono intrinsecamente stocastici: le stesse reti producono risposte leggermente diverse a stimoli identici, e percorsi multipli di elaborazione convergono alla medesima risposta comportamentale. La struttura omotopica delle ∞-categorie è il linguaggio matematico che questa indeterminatezza biologica richiede — non un difetto da correggere verso l'esattezza categoriale, ma una proprietà strutturale che la formalizzazione deve rispettare e non eliminare.

Il **teorema di riconoscimento di Lurie** afferma che le quasi-categorie sono esattamente le strutture presentabili come localizzazioni di categorie ordinarie rispetto

a certi morfismi (le equivalenze deboli), nella stessa misura in cui la teoria dell'omotopia si ottiene dalla categoria dei CW-complessi localizzando rispetto alle equivalenze omotopiche deboli. Applicato alla cognizione: la struttura omotopica dell'esperienza temporale è la localizzazione della struttura cognitiva esatta rispetto alle equivalenze cognitive deboli — transizioni inferenziali che portano allo stesso stato a meno di differenze cognitivamente irrilevanti. Questa è, nel linguaggio più preciso disponibile, la struttura formale di quella **sintesi cognitiva** che Kant denominava unità dell'appercezione trascendentale: non un atto rigido e deterministico di connessione di rappresentazioni, ma una struttura omotopica che connette stati attraverso una gerarchia infinita di equivalenze coerenti, la cui coerenza globale è garantita dalla condizione del corno interno nel senso di Joyal.

XXV. Il classifying space B𝔐 come topologia del soggetto

La realizzazione geometrica |N𝔐| — lo **spazio classificante** B𝔐 della categoria degli stati mentali — è lo spazio topologico che codifica la struttura omotopica dell'intera vita cognitiva di un soggetto. I suoi invarianti — i gruppi di omotopia π (B𝔐) e i gruppi di (co)omologia H_n(B𝔐, ℤ) — sono invarianti della struttura cognitiva: non cambiano sotto deformazioni continue delle transizioni, e catturano soltanto le proprietà di scala larga della mente.

Il gruppo fondamentale π_1(B𝔐) codifica i **cicli cognitivi**: sequenze di stati che ritornano al punto di partenza, strutture ricorsive del pensiero, sistemi di credenze autoreferenziali. Un gruppo fondamentale triviale indica che ogni ciclo cognitivo è omotopicamente banale — che ogni ragionamento ricorsivo si può deformare continuamente verso il ragionamento nullo, senza struttura topologica residua. Un gruppo fondamentale non triviale indica la presenza di strutture di pensiero che resistono alla destrutturazione: pensieri con una topologia intrinseca irriducibile, analoga ai nodi nella teoria dei nodi.

Il secondo gruppo di omotopia π_2(B𝔐) codifica le strutture di coerenza di ordine superiore: le relazioni tra cicli cognitivi distinti, i modi in cui sistemi di credenze diversi si intrecciano in una struttura bidimensionale di dipendenze reciproche. E i gruppi superiori π stratificano questa gerarchia a ogni profondità.

Questa prospettiva suggerisce un programma — la **topologia cognitiva** come studio degli invarianti omotopici dello spazio classificante del sistema cognitivo di un soggetto o di una comunità epistemica. Due soggetti cognitivamente equivalenti avrebbero B𝔐 omotopicamente equivalenti; la variabilità inter-individuale si esprimerebbe nella variabilità degli invarianti topologici. Soggetti con sistemi di credenze più rigidi — meno deformabili sotto pressione argomentativa — avrebbero gruppi di omotopia più ricchi; soggetti cognitivamente più flessibili avrebbero spazi classificanti topologicamente più semplici, avvicinandosi al caso contrattile (π = 0 per ogni n) che corrisponde a un sistema di credenze privo di struttura ricorsiva irriducibile.

Va riconosciuto, per chiudere questo arco argomentativo con onestà, che il programma della topologia cognitiva rimane puramente teorico. Non esistono metodi sperimentali per misurare gli invarianti topologici di $B\mathfrak{M}$ per un sistema cognitivo reale: la costruzione di $N\mathfrak{M}$ richiederebbe la conoscenza dell'intera struttura delle transizioni inferenziali del soggetto, che nessuno strumento di neuroimmagine o psicometria attuale è in grado di fornire. Ma l'impossibilità pratica non è impossibilità concettuale. Il valore dell'analisi risiede nella precisione con cui essa indica che cosa bisognerebbe misurare e quale struttura matematica ci si aspetta di trovare — una mappa dell'ignoto che, senza indicare la destinazione, precisa almeno la forma del percorso.

Appendice A — Elementi di teoria delle categorie monoïdali chiuse

A.1 Coerenza di Mac Lane e il teorema di coerenza

Le equazioni del pentagonale e del triangolare che una categoria monoïdale deve soddisfare non esauriscono, in linea di principio, l'insieme di tutti i diagrammi di coerenza che si potrebbe richiedere. Il **teorema di coerenza di Mac Lane** (1963) afferma tuttavia che esse sono sufficienti: ogni diagramma costruito a partire da associatori, unitatori e loro inversi, componendo con funtori $\otimes$ e con morfismi della categoria, commuta automaticamente. In modo più preciso, ogni due morfismi in una categoria monoïdale libera che abbiano lo stesso dominio e lo stesso codominio sono necessariamente uguali. Il teorema di coerenza permette dunque di lavorare «come se» la categoria monoïdale fosse strettamente associativa — come se $(A \otimes B) \otimes C = A \otimes (B \otimes C)$ fosse un'uguaglianza e non soltanto un isomorfismo — senza perdere generalità, a patto di non richiedere uguaglianze tra morfismi ma soltanto uguaglianze a meno di isomorfismo coerente.

La dimostrazione del teorema di coerenza si articola in due passi. Nel primo si costruisce la **categoria libera** monoïdale $\mathscr{C}$_lib generata da un insieme di oggetti generatori: i suoi oggetti sono tutte le espressioni parenteticizzate ottenute combinando i generatori con $\otimes$ e con I, e i suoi morfismi sono formalmente le composizioni di associatori e unitatori necessarie per passare da un'espressione parenteticizzata a un'altra. Nel secondo si mostra che in $\mathscr{C}$_lib ogni due morfismi paralleli — con lo stesso dominio e codominio — sono necessariamente uguali, usando la tecnica della riduzione di Mac Lane verso una «forma normale» che a ogni espressione parenteticizzata associa canonicamente la parenteticizzazione completamente a sinistra. La commutatività di tutti i diagrammi segue dall'unicità della forma normale.

A.2 Aggiunzioni e loro proprietà universali

Siano $\mathscr{C}$ e $\mathscr{D}$ due categorie e L: $\mathscr{D} \to \mathscr{C}$, R: $\mathscr{C} \to \mathscr{D}$ due funtori. Sono equivalenti le seguenti condizioni:

(i) Esiste una biiezione naturale $\varphi_{A,B}$: $\mathrm{Hom}_{\mathscr{C}}(L(B), A) \cong \mathrm{Hom}_{\mathscr{D}}(B, R(A))$ per ogni $A \in \mathscr{C}$, $B \in \mathscr{D}$.

(ii) Esistono trasformazioni naturali η: $\mathrm{Id}_{\mathscr{D}} \Rightarrow R \circ L$ e ε: $L \circ R \Rightarrow \mathrm{Id}_{\mathscr{C}}$ che soddisfano le identità triangolari:

$(\varepsilon * \mathrm{id}_L) \circ (\mathrm{id}_L * \eta) = \mathrm{id}_L$
$(\mathrm{id}_R * \varepsilon) \circ (\eta * \mathrm{id}_R) = \mathrm{id}_R$

dove $*$ indica la composizione orizzontale di trasformazioni naturali.

(iii) Per ogni $B \in \mathscr{D}$, la coppia $(L(B), \eta_B: B \to R(L(B)))$ è un oggetto universale iniziale nella categoria $(B \downarrow R)$ degli oggetti sotto R: per ogni $A \in \mathscr{C}$ e ogni morfismo $f: B \to R(A)$ esiste un unico morfismo $\bar{f}: L(B) \to A$ in $\mathscr{C}$ tale che $R(\bar{f}) \circ \eta_B = f$.

La corrispondenza $\varphi_{A,B}$ si ricostruisce a partire dall'unità come:

$\varphi_{A,B}(g: L(B) \to A) = R(g) \circ \eta_B: B \to R(A)$

e la sua inversa come:

$\varphi^{-1}_{A,B}(h: B \to R(A)) = \varepsilon_A \circ L(h): L(B) \to A$

La naturalità di φ in A significa che per ogni $f: A \to A'$ in $\mathscr{C}$ si ha:

$\varphi_{A',B}(f \circ g) = R(f) \circ \varphi_{A,B}(g)$

e la naturalità in B significa che per ogni $k: B' \to B$ in $\mathscr{D}$:

$\varphi_{A,B'}(g \circ L(k)) = \varphi_{A,B}(g) \circ k$

Le aggiunzioni si compongono: se $L_1 \dashv R_1$ tra $\mathscr{D}$ e $\mathscr{C}$ e $L_2 \dashv R_2$ tra $\mathscr{C}$ e $\mathscr{B}$, allora $L_2 \circ L_1 \dashv R_1 \circ R_2$ tra $\mathscr{D}$ e $\mathscr{B}$. I funtori aggiunti destri preservano i limiti: se $R: \mathscr{C} \to \mathscr{D}$ ha un aggiunto sinistro, allora R preserva tutti i limiti che esistono in $\mathscr{C}$. Dualmente, i funtori aggiunti sinistri preservano i colimiti.

A.3 Limiti e colimiti: costruzioni universali fondamentali

Sia $F: \mathscr{J} \to \mathscr{C}$ un funtore. Il **limite** $\lim F$, se esiste, è caratterizzato dalla proprietà universale: esiste un cono $\lambda = \{\lambda_j: \lim F \to F(j)\}_{j \in \mathscr{J}}$ tale che per ogni altro cono (L, μ) esiste un unico morfismo $m: L \to \lim F$ con $\lambda_j \circ m = \mu_j$ per ogni j.

Casi particolari di limiti:

— **Prodotto** di A e B: limite del diagramma discreto {A, B} (categoria indice senza morfismi non identità). Oggetto $A \times B$ con proiezioni $\pi_A: A \times B \to A$ e $\pi_B: A \times B \to B$ tali che per ogni Z con morfismi $f: Z \to A$ e $g: Z \to B$ esiste un unico $\langle f, g \rangle: Z \to A \times B$ con $\pi_A \circ \langle f, g \rangle = f$ e $\pi_B \circ \langle f, g \rangle = g$.

— **Equalizzatore** di $f, g: A \to B$: limite del diagramma parallelo $A \rightrightarrows B$. Oggetto Eq(f,g) con morfismo $e: Eq(f,g) \to A$ tale che $f \circ e = g \circ e$, universale con questa proprietà.

— **Pullback** di $f: A \to C$ e $g: B \to C$: limite del diagramma $A \to^f C \leftarrow^g B$. Oggetto $A \times_C B$ con morfismi $p: A \times_C B \to A$ e $q: A \times_C B \to B$ tali che $f \circ p = g \circ q$, con la proprietà universale: per ogni Z con $h: Z \to A$ e $k: Z \to B$ con $f \circ h = g \circ k$ esiste un unico $Z \to A \times_C B$ facente commutare il diagramma.

I colimiti sono le costruzioni duali:

— **Coprodotto** $A \sqcup B$: colimite del diagramma discreto {A, B}. — **Coequalizzatore** di $f, g: A \to B$: colimite del diagramma parallelo. — **Pushout** di $f: C \to A$ e $g: C \to B$: colimite del diagramma $A \leftarrow^f C \to^g B$, che «incolla» A e B lungo C.

In **Set** tutti i limiti e i colimiti esistono. Il pullback di $f: A \to C$ e $g: B \to C$ è l'insieme:

$$A \times_C B = \{ (a, b) \in A \times B : f(a) = g(b) \}$$

Il pushout di $f: C \to A$ e $g: C \to B$ è il quoziente:

$$A \sqcup_C B = (A \sqcup B) / \sim$$

dove $\sim$ è la relazione di equivalenza generata da $f(c) \sim g(c)$ per ogni $c \in C$.

Appendice E — Caso Modello: Varietà Rappresentazionale delle Cellule di Luogo Ippocampali

---

E.1 — Motivazione e scelta del sistema modello

Il sistema delle cellule di luogo ippocampali del roditore costituisce, allo stato attuale delle neuroscienze, il caso in cui la corrispondenza tra struttura geometrica dello spazio fisico e struttura geometrica dello spazio rappresentazionale è stata

caratterizzata con la maggiore precisione sperimentale e teorica. È dunque il banco di prova ideale per la teoria proposta nel corpo principale: se la metrica di Fisher intrinseca della varietà rappresentazionale coincide — o diverge in modo misurabile e interpretabile — dalla geometria dello spazio ambientale che le cellule di luogo codificano, si ottiene un test diretto del principio di individuazione del contenuto enunciato in IV.1.

Una cellula di luogo è un neurone ippocampale che scarica selettivamente quando l'animale si trova in una regione specifica dell'ambiente — il suo *campo di luogo* — e rimane sostanzialmente silente altrove. O'Keefe e Dostrovsky scoprirono questo fenomeno nel 1971; O'Keefe e Nadel lo svilupparono nella teoria della *cognitive map* del 1978, sostenendo che l'ippocampo implementa una mappa spaziale allocentrica dell'ambiente. Decades di ricerche successive hanno confermato e raffinato questa visione, culminando con il Nobel del 2014 assegnato a O'Keefe, Moser e Moser per la scoperta delle cellule di griglia, delle cellule di confine e delle cellule di testa.

E.2 — Il modello matematico delle cellule di luogo

Si consideri un ambiente bidimensionale convesso $\mathcal{A} \subset \mathbb{R}^2$, per esempio un quadrato di lato L. Una popolazione di N cellule di luogo è caratterizzata da N campi di luogo con centri $x_i \in \mathcal{A}$ e larghezze $\sigma_i > 0$. Il tasso di scarica medio della cellula i quando l'animale si trova nella posizione $x \in \mathcal{A}$ è modellato da una funzione di Von Mises (per ambienti toroidali) o, nella approssimazione gaussiana valida per ambienti non periodici, da:

$$r_i(x) = r_{max} \cdot \exp\left(-\|x - x_i\|^2 / (2\sigma_i^2) \right) + r_0$$

dove r_{max} è il tasso di picco, r_0 è il tasso basale, e $\|\cdot\|$ è la norma euclidea in $\mathbb{R}^2$. Il vettore di attivazione della popolazione è $r(x) = (r_1(x), \ldots, r_N(x)) \in \mathbb{R}^N$, e la mappa di codifica è $f: \mathcal{A} \to \mathbb{R}^N$, $f(x) = r(x)$.

Assumendo un modello di rumore di Poisson indipendente tra cellule — ogni cellula emette spikes secondo un processo di Poisson con tasso $r_i(x)$ — la distribuzione della risposta della popolazione a una posizione x è:

$$p(n_1, \ldots, n_N ; x) = \prod_i \left[r_i(x)^{n_i} \cdot e^{-r_i(x)} / n_i! \right]$$

dove n_i è il numero di spikes osservati dalla cellula i in un intervallo di tempo Δt (e $r_i(x)$ è ora il numero atteso di spikes, ossia il tasso moltiplicato per Δt).

La matrice di informazione di Fisher di questa distribuzione rispetto ai parametri $x = (x^1, x^2)$ — le coordinate spaziali — è:

$$g_{kl}(x) = \sum_i \left(1/r_i(x)\right) \cdot \left(\partial r_i / \partial x^k\right) \cdot \left(\partial r_i / \partial x^l\right)$$

dove si è trascurato il tasso basale r_0 nell'approssimazione in cui $r_0 \ll r_{max}$. Sostituendo la forma gaussiana:

$$\partial r_i / \partial x^k = r_i(x) \cdot (x_i^k - x^k) / \sigma_i^2$$

si ottiene:

$$g_{kl}(x) = \Sigma_i \, r_i(x) \cdot (x_i^k - x^k)(x_i^l - x^l) / \sigma_i^4$$

Questa è la metrica di Fisher della varietà rappresentazionale delle cellule di luogo, espressa nelle coordinate spaziali x. Essa è una matrice 2×2 semidefinita positiva che varia punto per punto nell'ambiente: misura quanto facilmente la posizione x può essere discriminata dalle posizioni vicine sulla base delle risposte della popolazione.

E.3 — Calcolo esplicito della metrica nel caso omogeneo

Si considera ora il caso semplificato in cui tutti i campi di luogo hanno la stessa larghezza $\sigma_i = \sigma$ e i centri x_i sono distribuiti uniformemente sull'ambiente con densità ϱ (numero di cellule per unità di area). In questo limite, la somma discreta si approssima con un integrale:

$$g_{kl}(x) \approx \varrho \int_{\mathcal{A}} r(x; x') \cdot (x'^k - x^k)(x'^l - x^l) / \sigma^4 \cdot d^2x'$$

dove $r(x; x') = r_{max} \cdot \exp(-\|x - x'\|^2/(2\sigma^2))$ è il tasso di scarica di una cellula centrata in x' quando l'animale è in x. Si introduce il cambio di variabile $u = x' - x$:

$$g_{kl}(x) \approx (\varrho \, r_{max} / \sigma^4) \int_{\mathbb{R}^2} \exp(-\|u\|^2/(2\sigma^2)) \cdot u^k \, u^l \cdot d^2u$$

L'integrale è gaussiano e si valuta esplicitamente. Per $k = l = 1$ (componente xx):

$$\int_{\mathbb{R}^2} \exp(-\|u\|^2/(2\sigma^2)) \cdot (u^1)^2 \cdot d^2u = \int \exp(-(u^1)^2/(2\sigma^2)) \, (u^1)^2 \, du^1 \cdot \int \exp(-(u^2)^2/(2\sigma^2)) \, du^2$$
$$= \sigma^3\sqrt{(2\pi)} \cdot \sigma\sqrt{(2\pi)} = 2\pi\sigma^4$$

Per $k \neq l$ (componente xy), l'integrale si annulla per simmetria. Per $k = l = 2$ si ottiene lo stesso risultato. Dunque:

$$g_{kl}(x) \approx (\varrho \, r_{max} / \sigma^4) \cdot 2\pi\sigma^4 \cdot \delta_{kl} = 2\pi\varrho \, r_{max} \cdot \delta_{kl}$$

Un risultato notevole: nel caso omogeneo (densità uniforme, larghezze uguali), la metrica di Fisher è *piatta e isotropa* — è proporzionale alla metrica euclidea di $\mathbb{R}^2$. La varietà rappresentazionale delle cellule di luogo, in questo limite ideale, è isometrica (a

meno di una costante) all'ambiente fisico. La curvatura è zero. Questo risultato è non triviale: significa che una popolazione di cellule di luogo uniformemente distribuite implementa una rappresentazione *fedele* dello spazio fisico nel senso geometrico più forte: la metrica del contenuto rappresentazionale coincide con la metrica dello spazio reale.

E.4 — Deviazioni dall'isometria: non-uniformità e struttura della curvatura

Il caso omogeneo è un ideale limite. Gli esperimenti mostrano sistematicamente deviazioni dall'uniformità: i campi di luogo sono distribuiti in modo non uniforme (con addensamenti vicino agli obiettivi di navigazione e agli ingressi), le larghezze σ_i variano in modo sistematico con la posizione, e le correlazioni tra cellule non sono trascurabili. Queste deviazioni producono una metrica di Fisher non piatta, e quindi una curvatura non nulla della varietà rappresentazionale.

Si considera ora il caso in cui la densità $\varrho(x)$ dei centri dei campi di luogo varia lentamente. Ripetendo il calcolo precedente con ϱ dipendente da x, si ottiene:

$$g_{kl}(x) \approx 2\pi\sigma^2 \, r_{max} \cdot \varrho(x) \cdot \delta_{kl} + O(\sigma^2 \, \partial\varrho)$$

La metrica è ancora diagonale al leading order, ma con un fattore conforme $\varrho(x)$ che varia nello spazio. Una metrica conforme piatta (del tipo $g_{kl} = \Omega(x) \, \delta_{kl}$) ha curvatura di Gauss:

$$K(x) = -(1/(2\Omega)) \, \Delta \log \Omega$$

dove Δ è il laplaciano euclideo e $\Omega(x) = 2\pi\sigma^2 \, r_{max} \cdot \varrho(x)$. Sostituendo:

$$\begin{aligned} K(x) &= -(1/(4\pi\sigma^2 \, r_{max} \cdot \varrho(x))) \cdot \Delta \log(\varrho(x)) \\ &= -(1/(4\pi\sigma^2 \, r_{max} \cdot \varrho(x))) \cdot [\, \Delta\varrho(x)/\varrho(x) - \| \nabla \varrho(x) \|^2/\varrho(x)^2 \,] \end{aligned}$$

Questa formula ha un'interpretazione diretta. Nelle regioni in cui la densità dei campi di luogo ha un massimo locale ($\Delta\varrho < 0$), la curvatura è positiva: la varietà rappresentazionale si incurva come una sfera in quelle regioni. Nelle regioni in cui la densità ha un minimo locale ($\Delta\varrho > 0$), la curvatura è negativa: la varietà si incurva come un iperboloide.

La connessione con la teoria del contenuto è immediata e falsificabile: le regioni dell'ambiente in cui i roitori trascorrono più tempo (e verso cui pertanto molte cellule sviluppano campi di luogo, producendo alta densità ϱ) sono anche le regioni in cui la varietà rappresentazionale ha curvatura positiva. Queste sono le regioni di maggiore «risoluzione» rappresentazionale, ma anche, secondo la teoria proposta, le regioni di contenuto più «compresso» e strutturato — l'analogo cognitivo di un punto focale. La predizione è che gli obiettivi di navigazione (le posizioni della ricompensa, i rifugi)

debbano corrispondere a regioni di curvatura positiva nella varietà neurale ippocampale.

E.5 — La struttura toroidale e la olonomia dello spazio periodico

Il risultato più rilevante per la teoria della olonomia sviluppata in V.4 riguarda gli ambienti periodici. Gardner e colleghi (2022, *Nature*) hanno dimostrato, usando tecniche di omologia persistente sulle registrazioni di cellule di luogo in ambiente toroidale (un labirinto rettangolare con condizioni al contorno periodiche), che la varietà neurale ha topologia toroidale $T^2 = S^1 \times S^1$, con i due cicli fondamentali del toro che corrispondono ai due assi del labirinto.

Si calcola ora esplicitamente l'olonomia della connessione di Levi-Civita per la varietà rappresentazionale in questo caso. Il toro piatto $T^2_{L_1, L_2}$ — il toro con lati L_1 e L_2, ottenuto identificando i lati opposti del rettangolo $[0, L_1] \times [0, L_2]$ — ha curvatura di Gauss $K = 0$ ovunque: è una varietà piatta. In una varietà piatta, il gruppo di olonomia è triviale per curve contraibili, ma non necessariamente per curve non contraibili.

Consideriamo il ciclo γ_1 che percorre il toro nella direzione x^1 (il loop orizzontale) con lunghezza L_1. Il trasporto parallelo di un vettore $v = (v^1, v^2)$ lungo γ_1 su un toro piatto è:

$$H_{\gamma_1}(v) = v \quad \text{(identità)}$$

Analogamente per γ_2. Dunque, il gruppo di olonomia del toro piatto è triviale: $\{Id\}$. Questo significa che un animale che naviga un ciclo completo dell'ambiente torna con il vettore di «orientazione cognitiva» identico a quello iniziale — non c'è accumulo di rotazione geometrica.

Il caso diventa più interessante quando si consideri il toro non piatto, prodotto da una densità non uniforme $\varrho(x)$. In questo caso la curvatura $K \neq 0$, e il teorema di Gauss-Bonnet sul toro afferma:

$$\int_{T^2} K \cdot dA = \chi(T^2) \cdot 2\pi = 0$$

poiché la caratteristica di Eulero del toro è $\chi(T^2) = 0$. Dunque la curvatura totale è nulla, ma la curvatura locale può variare: regioni positive bilanciano regioni negative. In questo caso, il trasporto parallelo lungo un loop γ che racchiude una regione di curvatura non nulla produce una rotazione H_γ di angolo:

$$\theta_\gamma = \iint_{\Sigma(\gamma)} K \cdot dA$$

dove $\Sigma(\gamma)$ è la regione delimitata da γ (usando il teorema di Gauss-Bonnet locale). Se l'animale naviga un percorso che racchiude una regione di alta densità di campi (alta curvatura positiva), il suo vettore di «orientazione rappresentazionale» torna ruotato di

un angolo proporzionale alla curvatura integrata racchiusa. Questa è la *olonomia spaziale* della navigazione: la memoria geometrica del percorso.

La predizione empirica è precisa: due percorsi che connettono gli stessi punti di partenza e arrivo ma racchiudono regioni di curvatura diversa produrranno rotazioni diverse degli stati cognitivi. Questa differenza è in linea di principio misurabile come differenza nella «fase» della risposta delle cellule di luogo — un effetto analogo ma distinto dal remapping ippocampale classico.

E.6 — Stima numerica e ordini di grandezza

Si fornisce ora una stima numerica dei parametri rilevanti per verificare la plausibilità quantitativa della teoria. Per il ratto che esplora un ambiente quadrato di lato L = 1 m, i valori tipici misurati sperimentalmente sono:

N ≈ 300-500 cellule di luogo registrate simultaneamente (frazione della popolazione totale)
$\sigma \approx 10\text{-}20$ cm (larghezza del campo di luogo)
$r_{max} \approx 5\text{-}20$ Hz (tasso di picco)
$\varrho \approx 300\text{-}500 / m^2$ (densità media dei centri, per la sottoregione CA1)

Con questi valori, il fattore conforme della metrica è:

$$\Omega = 2\pi\sigma^2 r_{max} \cdot \rho \approx 2\pi \cdot (0.15\ m)^2 \cdot 10\ Hz \cdot 400\ m^{-2} \approx 565\ m^{-2} \cdot (s^{-1} \cdot \Delta t)$$

dove Δt è la finestra temporale di integrazione. Per $\Delta t = 0.1$ s (tipica finestra per la decodifica di posizione):

$$\Omega \approx 56.5\ m^{-2}$$

La distanza di Fisher tra due posizioni x e x + dx è dunque:

$$ds = \sqrt{\Omega} \cdot \|dx\| \approx 7.5\ m^{-1} \cdot \|dx\|$$

In unità adimensionali (normalizzando la lunghezza con σ), la distanza di Fisher tra due posizioni separate di $\sigma = 15$ cm è:

$$ds_{Fisher} \approx 7.5 \cdot 0.15 \approx 1.1$$

ossia circa un'unità di «distanza discriminativa». Questo ha un'interpretazione statistica diretta: due posizioni separate di una larghezza di campo di luogo sono discriminabili con probabilità di errore ≈ 16% (corrispondente a una distanza di 1 sigma nella distribuzione gaussiana degli errori di decodifica ottimale). I dati sperimentali sulla

decodifica bayesiana della posizione mostrano errori di decodifica dell'ordine di 5-10 cm con popolazioni di 100-200 cellule, compatibile con questo ordine di grandezza.

E.7 — La curvatura come predizione testabile

Il risultato più direttamente testabile della teoria è la relazione tra densità dei campi di luogo e curvatura della varietà rappresentazionale. Si è mostrato in A.4 che:

$$K(x) \approx -\Delta \log(\rho(x)) / (4\pi\sigma^2 r_{max} \cdot \rho(x))$$

Questa formula permette di calcolare la curvatura della varietà dalla mappa di densità dei campi di luogo, che è misurabile sperimentalmente. La predizione è:

(*P1*) Nelle regioni in cui l'animale viene addestrato a trovare la ricompensa, la densità ρ aumenta (rimapping verso ricompense, documentato sperimentalmente da Dupret et al. 2010 e Hollup et al. 2001); dunque $K > 0$ in quelle regioni.

(*P2*) La curvatura K è misurabile indirettamente mediante omologia persistente: il diagramma di persistenza P_1 della varietà di attivazione mostra la nascita e la morte di buchi 1-dimensionali. In una varietà a curvatura prevalentemente positiva, le sfere geodetiche sono più piccole di quelle euclidee, e i buchi tendono a chiudersi più rapidamente nella filtrazione di Čech: il *lifetime* medio dei generators di H_1 è più corto nelle regioni di alta curvatura.

(*P3*) La rotazione holonomica H_γ prodotta da un percorso che racchiude una regione di alta densità è misurabile come sfasamento della risposta delle cellule di luogo prima e dopo la navigazione del percorso — analogo del *phase precession* ma su scala di loop, non di campo singolo.

Queste tre predizioni sono in linea di principio verificabili con tecniche sperimentali esistenti — registrazioni multi-tetrode o probe siliconiche ad alta densità (Neuropixels), combinate con analisi di omologia persistente sulla matrice di attivazione — e costituiscono quindi test empirici diretti della teoria geometrica del contenuto proposta in questa monografia.

Appendice B — Geometria dell'informazione: derivazioni complete

B.1 Derivazione della metrica di Fisher-Rao dalla divergenza di Kullback-Leibler

La metrica di Fisher-Rao emerge naturalmente dalla struttura locale della divergenza di Kullback-Leibler. Data la famiglia parametrica $\mathscr{S} = \{p(x \mid \theta)\}$, la divergenza KL tra due distribuzioni vicine $p(x \mid \theta)$ e $p(x \mid \theta + d\theta)$ è:

$$D_{KL}(p(\cdot|\theta) \,\|\, p(\cdot|\theta+d\theta)) = \int p(x|\theta) \log[\, p(x|\theta) \,/\, p(x|\theta+d\theta) \,]\, dx$$

Espandendo al second'ordine in $d\theta$:

$$\log p(x|\theta+d\theta) \approx \log p(x|\theta) + (\partial_i \log p(x|\theta))\, d\theta^i + (1/2)(\partial_i\partial_j \log p(x|\theta))\, d\theta^i\, d\theta^j + O(|d\theta|^3)$$

Sostituendo nell'integrale e usando che $\int p(x|\theta)\, dx = 1$, da cui:

$$\int p(x|\theta)\, \partial_i \log p(x|\theta)\, dx = \partial_i \int p(x|\theta)\, dx = 0$$

e quindi il termine di primo ordine scompare. Il termine di secondo ordine dà:

$$D_{KL}(p(\cdot|\theta) \,\|\, p(\cdot|\theta+d\theta)) \approx (1/2)\, g_{ij}(\theta)\, d\theta^i\, d\theta^j + O(|d\theta|^3)$$

dove:

$$\begin{aligned} g_{ij}(\theta) &= -\int p(x|\theta)\, \partial_i\partial_j \log p(x|\theta)\, dx \\ &= \int p(x|\theta)\, (\partial_i \log p(x|\theta))(\partial_j \log p(x|\theta))\, dx \\ &= E_{p(x|\theta)}[\, \partial_i \ell(\theta;x) \cdot \partial_j \ell(\theta;x)\,] \end{aligned}$$

dove $\ell(\theta;x) = \log p(x|\theta)$ è il log-likelihood. L'equivalenza tra le due forme si ottiene differenziando l'identità $\int p(x|\theta)\, \partial_i \log p(x|\theta)\, dx = 0$ rispetto a θ^j:

$$\int (\partial_j p(x|\theta))\, \partial_i \log p(x|\theta)\, dx + \int p(x|\theta)\, \partial_i\partial_j \log p(x|\theta)\, dx = 0$$
$$\int p(x|\theta)(\partial_j \log p)(\partial_i \log p)\, dx = -\int p(x|\theta)\, \partial_i\partial_j \log p\, dx$$

che è appunto l'uguaglianza tra le due forme di g_{ij}.

B.2 Geodetiche e distanza geodetica nello spazio statistico

Una **geodetica** sulla varietà statistica $(\mathcal{S}, g)$ è una curva $\gamma: [0,1] \to \mathcal{S}$ che minimizza la lunghezza:

$$L(\gamma) = \int_0^1 \sqrt{(\, g_{ij}(\gamma(t))\, \gamma^i(t)\, \gamma^j(t)\,)}\, dt$$

Le geodetiche soddisfano l'equazione delle geodetiche:

$$\gamma^k(t) + \Gamma^k_{ij}(\gamma(t))\, \gamma^i(t)\, \gamma^j(t) = 0$$

dove i simboli di Christoffel sono:

$$\Gamma^k_{ij} = (1/2)\, g^{kl}\, (\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij})$$

con g^{kl} inversa di g_{ij}. La **distanza geodetica** tra due distribuzioni $p_0 = p(\cdot|\theta_0)$ e $p_1 = p(\cdot|\theta_1)$ è:

$$d_{FR}(p_0, p_1) = \inf_{\gamma} L(\gamma)$$

dove l'inf è preso su tutte le curve γ che connettono θ_0 a θ_1. Per la famiglia gaussiana univariata $\mathcal{N}(\mu, \sigma^2)$ con $\theta = (\mu, \sigma)$, la metrica di Fisher-Rao è:

$$g = (1/\sigma^2)\, d\mu^2 + (2/\sigma^2)\, d\sigma^2$$

che è la metrica iperbolica sul semipiano superiore $\mathbb{H}^2 = \{(\mu,\sigma) : \sigma > 0\}$. La curvatura sezionale è costante $K = -1/2$, e la distanza geodetica è:

$$d_{FR}(\mathcal{N}(\mu_0,\sigma_0^2), \mathcal{N}(\mu_1,\sigma_1^2)) = \sqrt{2} \cdot \operatorname{arccosh}\left(1 + \frac{(\mu_1-\mu_0)^2 + 2(\sigma_1-\sigma_0)^2}{4\sigma_0\sigma_1}\right)$$

B.3 Connessioni duali di Amari e la geometria duale

La geometria dell'informazione di Amari introduce, accanto alla connessione di Levi-Civita ∇, una famiglia di connessioni parametrizzata da $\alpha \in \mathbb{R}$. La **connessione α di Amari** $\nabla^{(\alpha)}$ ha simboli di Christoffel:

$$\Gamma^{(\alpha)}_{ij,k} = \Gamma^{(0)}_{ij,k} - \frac{\alpha}{2} T_{ijk}$$

dove $\Gamma^{(0)}$ sono i simboli della connessione di Levi-Civita e T_{ijk} è il **tensore di skewness** (tensore di Amari):

$$T_{ijk}(\theta) = \mathbb{E}_{p(x|\theta)}\left[\partial_i \ell \cdot \partial_j \ell \cdot \partial_k \ell\right]$$

Le connessioni $\nabla^{(\alpha)}$ e $\nabla^{(-\alpha)}$ sono **duali** rispetto alla metrica g nel senso che per ogni tripletta di campi vettoriali X, Y, Z su $\mathcal{S}$:

$$X(g(Y, Z)) = g(\nabla^{(\alpha)}_X Y, Z) + g(Y, \nabla^{(-\alpha)}_X Z)$$

I casi $\alpha = 0$ (connessione di Levi-Civita), $\alpha = 1$ (connessione esponenziale o e-connessione) e $\alpha = -1$ (connessione mista o m-connessione) sono i più rilevanti. Una varietà statistica è **α-piatta** se la connessione $\nabla^{(\alpha)}$ ha curvatura nulla. Le famiglie esponenziali sono dualmente piatte: sono sia 1-piatte che (-1)-piatte, con i due sistemi di coordinate duali (η-coordinate naturali e θ-coordinate canonica). Su una varietà dualmente piatta, la **divergenza di Bregman** generalizza il concetto di distanza quadratica euclidea:

$$D_\varphi(\theta \,\|\, \theta') = \varphi(\theta) - \varphi(\theta') - (\theta - \theta')_i\, \partial^i \varphi(\theta')$$

dove φ è la funzione potenziale della coordinata duale, e soddisfa la **relazione di Pythagoras generalizzata**:

$$D_{\varphi}(\theta \,\|\, \theta'') = D_{\varphi}(\theta \,\|\, \theta') + D_{\varphi}(\theta' \,\|\, \theta'') + (\theta - \theta')_i \, (\partial_i \varphi(\theta'') - \partial_i \varphi(\theta'))$$

che per varietà ortogonali si riduce all'identità di Pitagora: la terza divergenza scompare e si ha $D_{\varphi}(\theta \,\|\, \theta'') = D_{\varphi}(\theta \,\|\, \theta') + D_{\varphi}(\theta' \,\|\, \theta'')$.

Appendice C — Topos: struttura interna e logica

C.1 Assiomi di un topos elementare

Un **topos elementare** (nel senso di Lawvere-Tierney) è una categoria $\mathscr{T}$ che soddisfa:

(T1) Limiti finiti. Per ogni piccola categoria indice finita $\mathscr{J}$ e ogni funtore $F: \mathscr{J} \to \mathscr{T}$, esiste $\lim F$ in $\mathscr{T}$. In particolare $\mathscr{T}$ ha oggetto terminale 1, prodotti $A \times B$, e pullback.

(T2) Classificatore di sottooggetti. Esiste un oggetto $\Omega \in \mathrm{Ob}(\mathscr{T})$ e un morfismo vero: $1 \to \Omega$ tale che per ogni morfismo mono $m: S \rightarrowtail A$ esiste un unico morfismo caratteristico $\chi_m: A \to \Omega$ che fa commutare il diagramma di pullback:

```
S ——————————> 1
|             |
m           vero
|             |
v             v
A ——χ_m——> Ω
```

e tale che ogni morfismo $A \to \Omega$ emerge come caratteristico di esattamente un sottooggetto di A.

(T3) Esponenziali. Per ogni coppia di oggetti $A, B \in \mathrm{Ob}(\mathscr{T})$ esiste un oggetto $B^A \in \mathrm{Ob}(\mathscr{T})$ (l'**oggetto funzione** da A a B) e un morfismo di valutazione $\mathrm{ev}: B^A \times A \to B$ tali che per ogni Z e ogni morfismo $f: Z \times A \to B$ esiste un unico $\tilde{f}: Z \to B^A$ con $\mathrm{ev} \circ (\tilde{f} \times \mathrm{id}_A) = f$.

In un topos si possono definire: le operazioni logiche $\wedge$, $\vee$, $\Rightarrow$, $\neg$ come morfismi $\Omega \times \Omega \to \Omega$ e $\Omega \to \Omega$; i quantificatori $\forall$ e $\exists$ come aggiunti al funtore di sostituzione; e si può sviluppare l'intera matematica elementare internamente al topos, usando la logica interna che risulta essere intuizionista in generale.

C.2 La logica interna di un topos e il suo significato cognitivo

La logica interna del topos $\mathscr{T}$ è costruita come segue. Gli **oggetti di tipo** sono gli oggetti di $\mathscr{T}$; le **formule** nel contesto di un oggetto X sono i morfismi $\varphi: X \to \Omega$; i

termini di tipo A nel contesto X sono i morfismi t: X → A. La congiunzione φ ∧ ψ di due formule φ, ψ: X → Ω è il morfismo:

X —⟨φ,ψ⟩→ Ω × Ω —∧→ Ω

dove ∧: Ω × Ω → Ω è il morfismo intersezione definito come il caratteristico del pullback di vero × vero lungo se stesso. La disgiunzione φ ∨ ψ è il caratteristico dell'immagine della mappa ⟨χ_φ, χ_ψ⟩: X → Ω × Ω nel suo codominio. La negazione ¬φ è il caratteristico del complemento del sottooggetto classificato da φ.

Il fatto cruciale è che la legge del terzo escluso φ ∨ ¬φ = vero non è un assioma della logica interna del topos: essa vale nel topos **Set** (dove Ω = {falso, vero} e la logica è classica) ma fallisce in generale. Nel topos dei fasci su uno spazio topologico X non discreto, Ω è il fascio degli aperti di X, e la verità di una formula è locale: una formula è vera nell'aperto U se e soltanto se è vera in ogni punto di U, ma può essere né vera né falsa nell'intero spazio X. Questa logica è esattamente quella di **Heyting**: la logica intuizionista in cui la verità è costruttiva e locale piuttosto che globale e bivalente.

Il significato cognitivo di questa struttura è che l'esperienza fenomenica — modellata come topos — ha una logica interna che non è necessariamente quella classica. Le proposizioni che descrivono i contenuti dell'esperienza possono trovarsi in stati di indeterminazione genuina che non si risolvono aggiungendo informazione: non «non so se φ è vero», ma «in questo contesto cognitivo, φ non è né vero né falso» — una forma di indeterminatezza costitutiva che la logica bivalente non può rappresentare senza distorsione.

C.3 Il lemma di Yoneda come universalità in un topos

In un topos $\mathscr{T}$, il funtore di Yoneda:

y: $\mathscr{T}$ → Fun($\mathscr{T}$^{op}, Set), A ↦ Hom_$\mathscr{T}$(-, A)

è pienamente fedele. L'oggetto classificante Ω gioca un ruolo speciale: per ogni oggetto A, il morfismo χ: Hom_$\mathscr{T}$(-, A) → Hom_$\mathscr{T}$(-, Ω^A) definito dalla trasposta dell'identità su Ω^A è una trasformazione naturale che «internalizza» i morfismi in A all'interno del topos. In questo senso Ω^A è l'**oggetto di potenze** di A — l'oggetto che classifica i sottooggetti di A × (-) — e il lemma di Yoneda interno al topos afferma che:

Hom_$\mathscr{T}$(X, Ω^A) ≅ Sub_$\mathscr{T}$(X × A)

naturalmente in X, dove Sub_$\mathscr{T}$(B) è l'insieme dei sottooggetti di B in $\mathscr{T}$. Nella lettura cognitiva: i contenuti intenzionali dell'oggetto A — i modi in cui A può essere «sottoposto» a un contesto X — sono completamente classificati dai morfismi X → Ω^A, ossia dall'oggetto di potenze di A rispetto al classificatore fenomenico Ω.

Appendice F — Teoria del Gauge e Composizionalità del Contenuto: Sviluppo Formale

F.1 — Fibrati principali e connessioni di Ehresmann: riepilogo

Nel corpo principale (sezione IV.6) si è introdotto il linguaggio dei fibrati di Ehresmann per rendere conto della composizionalità del contenuto mentale, con l'osservazione che la curvatura della connessione di fibrato è l'analogo cognitivo dell'intensità di un campo di gauge. Qui si sviluppa questa analogia con il rigore formale che le compete, mostrando che non si tratta di una metafora ornamentale ma di una struttura matematica che genera predizioni concrete sulla fenomenologia della composizione concettuale.

Un *fibrato principale* con gruppo strutturale G è una quadrupla (P, M, π, G) dove P è la *varietà totale*, M è la *varietà base*, π: P → M è la *proiezione* surjettiva, e G è un gruppo di Lie che agisce liberamente e transitivamente sulle fibre $\pi^{-1}(m)$ per ogni $m \in M$. Localmente, $P \cong M \times G$ (il fibrato è localmente triviale), ma globalmente la struttura può essere non triviale — e questa non-trivialità è precisamente la *topologia del gauge*.

Una *connessione di Ehresmann* su P è una decomposizione, differenziabile e G-equivariante, dello spazio tangente $T_p P$ in ogni punto $p \in P$ nella somma diretta di una *componente verticale* $V_p P$ (tangente alla fibra) e una *componente orizzontale* $H_p P$ (complemento scelto dalla connessione):

$$T_p P = V_p P \oplus H_p P$$

La componente verticale $V_p P$ è canonicamente definita: è il nucleo della differenziale $d\pi_p: T_p P \to T_{\pi(p)} M$. La componente orizzontale $H_p P$ è la scelta aggiuntiva — il «modo di trasportare orizzontalmente» attraverso il fibrato — ed è quello che costituisce la connessione.

Algebricamente, la connessione è equivalente a una *1-forma di connessione* $\mathcal{A} \in \Omega^1(P; \mathfrak{g})$ a valori nell'algebra di Lie $\mathfrak{g}$ di G, che soddisfa:

$\mathcal{A}(X^\#) = X$ per ogni $X \in \mathfrak{g}$ (dove $X^\#$ è il campo fondamentale su P)
$R_g^* \mathcal{A} = \mathrm{Ad}_{g^{-1}} \circ \mathcal{A}$ per ogni $g \in G$ (equivarianza sotto G)

La *curvatura* della connessione è la 2-forma a valori in $\mathfrak{g}$:

$$\mathcal{F} = d\mathcal{A} + (1/2)[\mathcal{A}, \mathcal{A}] \in \Omega^2(P; \mathfrak{g})$$

dove il secondo termine contiene il bracket nell'algebra di Lie. Per gruppi abeliani (G = U(1), come nell'elettromagnetismo), il bracket si annulla e $\mathcal{F} = d\mathcal{A}$ è semplicemente la derivata esterna della 1-forma di connessione. Per gruppi non abeliani (come SU(2) nella teoria di Yang-Mills), il termine quadratico è essenziale.

F.2 — Il fibrato della composizione concettuale

Si applica ora questa struttura alla composizione di contenuti mentali. Si consideri la situazione in cui un contenuto concettuale ha due componenti strutturalmente dipendenti — per esempio, il contenuto «rosso brillante» ha una componente cromatica (il tipo di colore) e una componente luminosa (l'intensità). La composizione di queste due componenti non è, come si è argomentato in IV.6, semplicemente il prodotto riemanniano delle varietà corrispondenti: le due componenti non sono indipendenti — la percezione di «brillante» non è identica per tutti i colori, e la percezione di «rosso» cambia in funzione dell'intensità (effetto Bezold-Brücke).

Si modella questa situazione come segue. La varietà base M è la varietà delle qualità cromatiche — lo spazio dei colori puri, identificabile con un sottoinsieme del piano proiettivo $\mathbb{R}P^2$ (il cerchio cromatico è S^1). La fibra tipica F è la varietà delle intensità luminose — identificabile con $\mathbb{R}_+$. Il fibrato totale P è lo spazio dei contenuti visivi colorati con intensità, e la proiezione $\pi: P \to M$ dimentica l'intensità.

Il gruppo strutturale G di questo fibrato è il gruppo delle trasformazioni che agisce sulle fibre preservando la struttura: poiché le fibre sono $\mathbb{R}_+$, le trasformazioni naturali sono le riscalature, $G = \mathbb{R}_+$ con la moltiplicazione — un gruppo di Lie abeliano unidimensionale. La 1-forma di connessione $\mathcal{A} \in \Omega^1(P; \mathbb{R})$ specifica come l'intensità «si trasporta orizzontalmente» al variare della qualità cromatica: come cambia la percezione di intensità quando si varia il colore mantenendo costante l'intensità fisica.

La curvatura $\mathcal{F} = d\mathcal{A} \in \Omega^2(M; \mathbb{R})$ misura il fallimento della commutatività del trasporto: se si varia prima il colore e poi l'intensità, o prima l'intensità e poi il colore, si arriva in generale a stati percettivi diversi. Formalmente, per due vettori tangenti $u, v \in T_m M$ nello spazio cromatico, la curvatura è:

$$\mathcal{F}(u, v) = u(\mathcal{A}(\tilde{v})) - v(\mathcal{A}(\tilde{u})) - \mathcal{A}([\tilde{u}, \tilde{v}])$$

dove $\tilde{u}, \tilde{v}$ sono i sollevamenti orizzontali di u, v. Il valore $\mathcal{F}(u, v)$ misura la «rotazione» dell'intensità prodotta da un loop infinitesimo nel piano cromatico con area dA(u, v): essa è diversa da zero se e solo se la percezione dell'intensità dipende dal colore in modo non triviale.

F.3 — L'effetto Bezold-Brücke come curvatura di gauge

L'*effetto Bezold-Brücke* è un fenomeno psicofisico ben documentato: all'aumentare della luminanza, la tonalità percepita di molti colori spettrali si sposta sistematicamente verso il blu o il giallo (gli spettri di 478 nm e 578 nm sono invarianti alla luminanza, tutti gli altri si spostano). Questo significa che la mappa «stessa intensità fisica, colore

diverso» produce effettivamente stati percettivi diversi — esattamente ciò che la curvatura non nulla della connessione predice.

Si può quantificare questa curvatura usando i dati psicofisici disponibili. Sia λ la lunghezza d'onda dello stimolo (la coordinata cromatica), e sia h(λ, L) la tonalità percepita (in gradi di cerchio cromatico) per una luce di lunghezza d'onda λ e luminanza L. L'effetto Bezold-Brücke afferma che $\partial h/\partial L \neq 0$ per la maggior parte dei λ. La 1-forma di connessione nel piano (λ, L) è:

$$\mathcal{A} = A_\lambda(\lambda, L)\, d\lambda + A_L(\lambda, L)\, dL$$

dove A_λ e A_L sono le componenti. La curvatura è:

$$\mathcal{F}_{\lambda L} = \partial A_L/\partial\lambda - \partial A_\lambda/\partial L$$

Dai dati di Purdy (1931) e dalle misure più recenti di Ayama, Nakatsue e Kaiser, la variazione di tonalità per lunghezze d'onda nella regione 490-560 nm (verde-giallo) all'aumentare della luminanza da 10 a 1000 troland è dell'ordine di 5-15°. Questo implica una curvatura non nulla della connessione nella regione mediana dello spazio cromatico, con un'intensità:

$$\mathcal{F}_{\lambda L} \approx \Delta h / (\Delta\lambda \cdot \Delta L) \approx 10° / (70\ \text{nm} \cdot 3\ \text{decades}) \approx 0.05\ °/(\text{nm} \cdot \text{decade})$$

Questo valore è piccolo ma misurabile, e fornisce una stima diretta della curvatura del fibrato percettivo colore-intensità. La predizione della teoria è che questa curvatura non sia arbitraria ma sia determinata dalla struttura delle connessioni neurali tra le aree cromatiche (V4, V8) e le aree di elaborazione della luminanza (V1, V2): la curvatura del fibrato cognitivo riflette la struttura dei pesi sinaptici tra queste aree.

F.4 — Invarianza di gauge e invarianti del contenuto

In fisica, l'invarianza di gauge è la proprietà che le equazioni del moto siano invarianti sotto trasformazioni locali del gruppo di gauge G: $g(x) \in G$ applicato indipendentemente in ogni punto x dello spazio-tempo. I campi fisicamente osservabili sono quelli che rimangono invarianti sotto queste trasformazioni — le grandezze che non dipendono dalla «scelta di gauge», ossia dalla scelta arbitraria della sezione del fibrato.

L'analogo cognitivo di questa struttura è il seguente. Si considerino le trasformazioni «interne» del sistema di rappresentazione: modificazioni dei valori assoluti di attivazione che non cambiano le relazioni relative tra attivazioni. Per esempio, una riscalatura omogenea $r(x) \to \lambda r(x)$ di tutte le attivazioni della stessa quantità non cambia le differenze relative tra cellule, e quindi non dovrebbe cambiare il contenuto rappresentato. Questo è analogo a una trasformazione di gauge globale: $g \to \text{const} \cdot g$.

Le trasformazioni di gauge *locali* — riscalature che variano punto per punto nello spazio dei contenuti — corrispondono a trasformazioni delle rappresentazioni che variano in modo spazialmente eterogeneo. L'invarianza sotto questi trasformazioni locali è una condizione molto più forte: essa richiede che il contenuto sia determinato solo dalle relazioni *relative* tra attivazioni in ogni intorno, non dai loro valori assoluti.

La 1-forma di connessione $\mathcal{A}$ — che nella fisica corrisponde al potenziale vettore elettromagnetico A_μ — è la struttura che permette di confrontare i valori delle attivazioni in punti diversi dello spazio dei contenuti in modo gauge-covariante. La curvatura $\mathcal{F}$ — il campo elettromagnetico $F_{\mu\nu}$ — è l'unico invariante di gauge che contiene informazione fisica non ridondante. Nella teoria del contenuto mentale, $\mathcal{F}$ è l'unico invariante della composizione che non dipende dalla parametrizzazione arbitraria delle componenti.

Questa struttura ha una conseguenza filosofica importante: l'identità di contenuto tra stati in sistemi diversi non dovrebbe essere giudicata confrontando i valori assoluti delle attivazioni (non gauge-invarianti) ma la struttura della curvatura $\mathcal{F}$ (gauge-invariante). Due sistemi — uno neuronale e uno artificiale — rappresentano lo stesso contenuto composto non se le loro attivazioni sono identiche, ma se le curvature dei loro fibrati percettivi sono isometriche.

F.5 — Anomalie di gauge e rottura della composizionalità

In fisica delle particelle, le *anomalie di gauge* sono configurazioni in cui la simmetria di gauge è soddisfatta a livello classico ma non sopravvive alla quantizzazione: i diagrammi di Feynman con loop fermionici producono termini che violano la conservazione della corrente di gauge. Le anomalie sono inaccettabili in una teoria fisica coerente e la loro cancellazione impone vincoli forti sul contenuto della teoria (per esempio, il numero di colori nella QCD).

Esiste un analogo cognitivo delle anomalie di gauge, che qui si propone per la prima volta. Si consideri la composizione di tre o più componenti di contenuto: per esempio, «rosso brillante caldo» con tre componenti — cromatica, luminosa e termica (temperatura del colore). Se ciascuna coppia di componenti ha una curvatura di fibrato non nulla, la composizione delle tre componenti dovrebbe essere descritta da un 2-fibrato (o gerba), una struttura di categoria superiore in cui i fibrati stessi formano le fibre.

Una *anomalia cognitiva di gauge* è una situazione in cui la composizione di tre componenti è internamente incoerente: la curvatura del 2-fibrato ha un'ostruzione topologica alla definizione di una connessione globale. Matematicamente, questa ostruzione è la *classe di Dixmier-Douady* $\mathrm{dd}(P) \in H^3(M; \mathbb{Z})$ del gerbo associato: se questa classe è non nulla, non esiste una sezione globale del gerbo, e la composizione delle tre componenti è topologicamente incoerente.

La predizione cognitiva è la seguente: i sistemi di rappresentazione che tentano di comporre tre o più componenti di contenuto mutuamente dipendenti potrebbero incontrare incoerenze sistematiche — casi in cui non è possibile attribuire un

contenuto composito ben definito perché la struttura di fibrato non ammette una sezione globale. Questi casi potrebbero corrispondere a certi tipi di incoerenza percettiva — illusioni «impossibili» come i colori di Land, o certi effetti di cross-modal incongruence — in cui il sistema percettivo non riesce a costruire un percetto composito stabile.

La classe di Dixmier-Douady è calcolabile (almeno in principio) dai dati psicofisici sulle curvature delle connessioni delle coppie di componenti, usando la formula:

$$dd(P) = [\mathscr{F}_{12} \wedge \mathscr{F}_{23} \wedge \mathscr{F}_{31}] \in H^3(M; \mathbb{R})$$

dove $\mathscr{F}\_\{ij\}$ è la curvatura della connessione tra le componenti i e j, e il prodotto wedge di tre 2-forme dà una 6-forma integrata sulla varietà tridimensionale delle terne di stati. Se questa quantità è non nulla (mod $\mathbb{Z}$ in coomologia intera), si prevede un'incoerenza sistematica nella composizione delle tre componenti.

Questo è un risultato genuinamente nuovo che la struttura di gauge del contenuto permette di derivare, e che non sarebbe derivabile da nessuna delle teorie del contenuto standard. Costituisce una predizione empirica non ovvia — la correlazione tra misure psicofisiche di curvatura percettiva e misure di incoerenza percettiva multi-componente — che potrebbe essere testata con protocolli psicofisici appropriati.

F.6 — Invarianti topologici del gauge e numeri quantici del contenuto

Si conclude questa appendice con un'osservazione sulla struttura topologica globale dei fibrati del contenuto. I fibrati principali su una varietà base M sono classificati da un gruppo di coomologia: per G = U(1), dalla coomologia intera $H^2(M; \mathbb{Z})$ (i cui generatori sono le *classi di Chern* c_1); per G = SU(2), da $H^4(M; \mathbb{Z})$ (le classi di Pontryagin p_1). Questi invarianti interi — i *numeri di Chern* e i *numeri di Pontryagin* — classificano i tipi qualitativamente distinti di composizione del contenuto.

Per la varietà cromatica $M = S^2$ (la sfera dei colori), $H^2(S^2; \mathbb{Z}) = \mathbb{Z}$, e il numero di Chern $c_1 \in \mathbb{Z}$ classifica i fibrati U(1) su S^2 (per esempio il fibrato di Hopf ha $c_1 = 1$). Nella teoria del contenuto, c_1 misura il numero di «avvolgimenti» del fibrato di intensità attorno alla sfera cromatica: un valore $c_1 = 0$ corrisponde a un fibrato triviale (intensità e cromatica indipendenti), un valore $c_1 \neq 0$ corrisponde a una dipendenza topologicamente non triviale che non può essere eliminata da nessuna scelta di parametrizzazione locale.

I numeri di Chern del fibrato del contenuto sono, in questo senso, i *numeri quantici del contenuto stesso*: interi che classificano qualitativamente i diversi tipi di composizione semantica e che non variano per deformazioni continue della connessione. Essi sono la firma topologica più robusta della struttura del contenuto composto, e la loro misura — tramite integrazione della curvatura su cicli appropriati della varietà cromatica:

$c_1 = (1/2\pi) \int_{S^2} \mathscr{F}$

costituisce un test della struttura topologica globale del contenuto percettivo che è indipendente dai dettagli metrici e pertanto particolarmente robusto rispetto al rumore sperimentale.

Appendice D — Quasi-categorie: la condizione del corno e l'omotopia superiore

D.1 Definizione formale di quasi-categoria

Il **corno** Λ^n per $0 \leq k \leq n$ è il sottinsieme simpliciale di Δ^n generato da tutte le facce tranne la k-esima:

$$\Lambda^n = \bigcup_{i \neq k} d_i(\Delta^{n-1})$$

Il corno Λ^n_0 e Λ^n sono **corni esterni** (rispettivamente sinistro e destro); i corni Λ^n per $0 < k < n$ sono **corni interni**.

Un insieme simpliciale X è un **Kan complex** se ogni corno (interno ed esterno) ammette un riempimento:

$$\forall\, n \geq 1,\ \forall\, k,\ \forall\, (\Lambda^n \to X)\ \exists\, (\Delta^n \to X) \text{ estensione}$$

Un insieme simpliciale X è una **quasi-categoria** (nel senso di Joyal) se ogni corno **interno** ammette un riempimento:

$$\forall\, n \geq 2,\ \forall\, 0 < k < n,\ \forall\, (\Lambda^n \to X)\ \exists\, (\Delta^n \to X) \text{ estensione}$$

La differenza è significativa: i Kan complex modellano gli ∞-groupoid (in cui tutti i morfismi sono invertibili), mentre le quasi-categorie modellano le ∞-categorie (in cui i morfismi non sono necessariamente invertibili, ma la composizione è definita a meno di omotopia).

Il nervo $N\mathscr{C}$ di una categoria ordinaria $\mathscr{C}$ è una quasi-categoria in cui i riempimenti sono unici (non soltanto esistenti), e questa unicità equivale alla condizione di Segal. Per una quasi-categoria generale, i riempimenti non sono unici, e le diverse scelte di riempimento sono collegate da strutture di ordine superiore.

D.2 Morfismi, equivalenze e il ∞-groupoid fondamentale

In una quasi-categoria X, gli oggetti sono gli 0-simplessi X_0 e i morfismi sono gli 1-simplessi X_1, con dominio e codominio:

$$\mathrm{dom}(f) = d_1(f) \in X_0, \quad \mathrm{cod}(f) = d_0(f) \in X_0 \quad \text{per } f \in X_1$$

La composizione di f: A → B e g: B → C è un morfismo h: A → C definito come il riempimento del corno interno Λ^2_1:

$\Lambda^2_1 = \{ d_0 \mapsto g, d_2 \mapsto f \} \rightarrow \Delta^2$

e h = d_1 del 2-simplesso riempiente. Il riempimento non è unico, ma due scelte diverse h e h' sono **omotope**: esiste un 3-simplesso che le connette.

Un morfismo f ∈ X_1 è un'**equivalenza** in X se esiste g: B → A tale che g ∘ f ≃ id_A e f ∘ g ≃ id_B, dove ≃ denota omotopia nel senso dei cammini nel ∞-groupoid sottostante. Il **∞-groupoid fondamentale** di X è il Kan complex ottenuto invertendo formalmente tutte le equivalenze in X — la localizzazione della quasi-categoria rispetto alle equivalenze, che coincide con la realizzazione geometrica |X| come spazio topologico.

D.3 Il teorema di Joyal e l'equivalenza con le categorie topologiche

Il **teorema di Joyal** (2002, con rafforzamenti di Lurie nel 2009) stabilisce l'equivalenza, in un senso preciso di equivalenza di teorie dell'omotopia, tra le seguenti strutture:

(a) Quasi-categorie (insiemi simpliciali con la condizione del corno interno). (b) Categorie topologiche (categorie arricchite su spazi topologici con la topologia della compatta-apertura). (c) Categorie simpliciali (categorie arricchite su insiemi simpliciali). (d) Categorie di Segal (prefasci Segal-enriched su Δ che soddisfano una condizione di completezza).

Questa equivalenza — dimostrata attraverso una catena di equivalenze di categorie di modelli nella struttura della teoria dell'omotopia di Quillen — garantisce che il concetto di ∞-categoria sia robusto: indipendente dal modello scelto per formalizzarlo. Per la teoria della cognizione che si è sviluppata nel testo, questa robustezza è rilevante: la struttura omotopica dell'esperienza temporale non dipende dalla scelta di un particolare modello di ∞-categoria, ma è un invariante intrinseco che qualsiasi modello adeguato deve catturare.

Bibliografia

Opere filosofiche

Benacerraf, P. (1965). What numbers could not be. *The Philosophical Review*, 74(1), 47–73.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9, 261–325.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.

Brandom, R. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge MA.

Brandom, R. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge MA.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Duncker & Humblot, Leipzig. [Trad. it.: *La psicologia dal punto di vista empirico*, Laterza, Bari, 1997.]

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, New York.

Davidson, D. (1970). Mental events. In L. Foster & J. Swanson (a cura di), *Experience and Theory*. University of Massachusetts Press. [Ripubblicato in *Essays on Actions and Events*, Oxford University Press, 1980.]

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company, Boston.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50. [Trad. it.: Senso e denotazione, in *Logica e aritmetica*, Boringhieri, Torino, 1965.]

Husserl, E. (1928). *Vorlesungen zur Phänomenologie des inneren Zeitbewusstseins*. Herausgegeben von Martin Heidegger. Max Niemeyer Verlag, Halle. [Trad. it.: *Per la fenomenologia della coscienza interna del tempo*, Franco Angeli, Milano, 1981.]

Husserl, E. (1950). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Erstes Buch*. Martinus Nijhoff, Den Haag. [Husserliana III.]

Kaplan, D. (1989). Demonstratives. In J. Almog, J. Perry & H. Wettstein (a cura di), *Themes from Kaplan*. Oxford University Press, New York, 481–563.

Kim, J. (1984). Concepts of supervenience. *Philosophy and Phenomenological Research*, 45(2), 153–176.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press, Cambridge.

Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Gallimard, Paris. [Trad. it.: *Fenomenologia della percezione*, Il Saggiatore, Milano, 1965.]

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press, Cambridge MA.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (a cura di), *Art, Mind, and Religion*. University of Pittsburgh Press, 37–48.

Russell, B. (1903). *The Principles of Mathematics*. Cambridge University Press, Cambridge.

Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, 79(7), 357–381.

Teoria delle categorie e matematica

Awodey, S. (2010). *Category Theory*. Second edition. Oxford University Press, Oxford.

Barr, M. & Wells, C. (1990). *Category Theory for Computing Science*. Prentice Hall, Englewood Cliffs NJ.

Eilenberg, S. & Mac Lane, S. (1945). General theory of natural equivalences. *Transactions of the American Mathematical Society*, 58(2), 231–294.

Eilenberg, S. & Kelly, G. M. (1966). Closed categories. In *Proceedings of the Conference on Categorical Algebra (La Jolla 1965)*. Springer, New York, 421–562.

Gabriel, P. & Zisman, M. (1967). *Calculus of Fractions and Homotopy Theory*. Springer, Berlin.

Grothendieck, A. (1957). Sur quelques points d'algèbre homologique. *Tôhoku Mathematical Journal*, 9(2), 119–221.

Joyal, A. (2002). Quasi-categories and Kan complexes. *Journal of Pure and Applied Algebra*, 175(1–3), 207–222.

Kan, D. M. (1958). Adjoint functors. *Transactions of the American Mathematical Society*, 87(2), 294–329.

Kelly, G. M. (1982). *Basic Concepts of Enriched Category Theory*. London Mathematical Society Lecture Note Series, vol. 64. Cambridge University Press, Cambridge. [Ristampato in: *Reprints in Theory and Applications of Categories*, 10 (2005).]

Lawvere, F. W. (1969). Adjointness in foundations. *Dialectica*, 23(3–4), 281–296.

Lawvere, F. W. (1973). Metric spaces, generalized logic, and closed categories. *Rendiconti del Seminario Matematico e Fisico di Milano*, 43(1), 135–166. [Ristampato in: *Reprints in Theory and Applications of Categories*, 1 (2002).]

Lawvere, F. W. & Tierney, M. (1970). Quantifiers and sheaves. In *Actes du Congrès International des Mathématiciens (Nice 1970)*, vol. 1. Gauthier-Villars, Paris, 329–334.

Lurie, J. (2009). *Higher Topos Theory*. Annals of Mathematics Studies, vol. 170. Princeton University Press, Princeton NJ.

Lurie, J. (2017). *Higher Algebra*. Disponibile all'indirizzo: https://www.math.ias.edu/~lurie/papers/HA.pdf

Mac Lane, S. (1963). Natural associativity and commutativity. *Rice University Studies*, 49(4), 28–46.

Mac Lane, S. (1971). *Categories for the Working Mathematician*. Graduate Texts in Mathematics, vol. 5. Springer, New York. [Seconda edizione: 1998.]

Mac Lane, S. & Moerdijk, I. (1992). *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Universitext. Springer, New York.

Riehl, E. (2014). *Categorical Homotopy Theory*. New Mathematical Monographs, vol. 24. Cambridge University Press, Cambridge.

Riehl, E. (2016). *Category Theory in Context*. Dover Publications, Mineola NY.

Segal, G. (1974). Categories and cohomology theories. *Topology*, 13(3), 293–312.

Geometria dell'informazione e statistica

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, vol. 28. Springer, Berlin.

Amari, S. (2016). *Information Geometry and Its Applications*. Applied Mathematical Sciences, vol. 194. Springer, Tokyo.

Amari, S. & Nagaoka, H. (2000). *Methods of Information Geometry*. Translations of Mathematical Monographs, vol. 191. American Mathematical Society & Oxford University Press, Providence RI.

Ay, N., Jost, J., Lê, H. V. & Schwachhöfer, L. (2017). *Information Geometry*. Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 64. Springer, Cham.

Chentsov, N. N. (1972). *Statistical Decision Rules and Optimal Inference*. Nauka, Mosca. [Trad. ingl.: Translations of Mathematical Monographs, vol. 53, American Mathematical Society, 1982.]

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5), 700–725.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3), 81–91.

Neuroscienze computazionali e teoria predittiva

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press, New York.

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K. J., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology — Paris*, 100(1–3), 70–87.

Helmholtz, H. von (1867). *Handbuch der physiologischen Optik*. Voss, Leipzig.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12), 712–719.

Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79–87.

Teoria dell'informazione integrata

Oizumi, M., Albantakis, L. & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, 10(5), e1003588.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biological Bulletin*, 215(3), 216–242.

Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.

Probabilità categoriale e categorie di Markov

Cho, K. & Jacobs, B. (2019). Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7), 938–971.

Fritz, T. (2020). A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370, 107239.

Giry, M. (1982). A categorical approach to probability theory. In *Categorical Aspects of Topology and Analysis*. Lecture Notes in Mathematics, vol. 915. Springer, Berlin, 68–85.

Panangaden, P. (1999). The category of Markov kernels. *Electronic Notes in Theoretical Computer Science*, 22, 171–187.

Topologia algebrica e teoria dell'omotopia

Hatcher, A. (2002). *Algebraic Topology*. Cambridge University Press, Cambridge. [Liberamente disponibile all'indirizzo: https://pi.math.cornell.edu/~hatcher/AT/ATpage.html]

May, J. P. (1967). *Simplicial Objects in Algebraic Topology*. Van Nostrand Mathematical Studies, vol. 11. D. Van Nostrand, Princeton NJ.

Quillen, D. G. (1967). *Homotopical Algebra*. Lecture Notes in Mathematics, vol. 43. Springer, Berlin.

Weibel, C. A. (1994). *An Introduction to Homological Algebra*. Cambridge Studies in Advanced Mathematics, vol. 38. Cambridge University Press, Cambridge.

Neuroscienze computazionali e varietà neurali

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., e DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology*, 10(12), e1003963.

Chung, S., Lee, D. D., e Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3), 031003.

Crick, F., e Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, 2, 263–275.

Curto, C., e Itskov, V. (2008). Cell groups reveal structure of stimulus space. *PLOS Computational Biology*, 4(10), e1000205.

DiCarlo, J. J., e Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8), 333–341.

Dupret, D., O'Neill, J., Pleydell-Bouverie, B., e Bhattacharya, S. (2010). The reorganization and reactivation of hippocampal maps predict spatial memory performance. *Nature Neuroscience*, 13(8), 995–1002.

Gardner, R. J., Hermansen, E., Pachitariu, M., Burak, Y., Baas, N. A., Dunn, B. A., Moser, M.-B., e Moser, E. I. (2022). Toroidal topology of population activity in grid cells. *Nature*, 602, 123–128.

Hollup, S. A., Molden, S., Donnett, J. G., Moser, M.-B., e Moser, E. I. (2001). Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *Journal of Neuroscience*, 21(5), 1635–1644.

Levina, E., e Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17* (NIPS 2004), 777–784. MIT Press, Cambridge (MA).

O'Keefe, J., e Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1), 171–175.

O'Keefe, J., e Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford.

Shenoy, K. V., Sahani, M., e Churchland, M. M. (2013). Cortical control of arm movements: A dynamical systems perspective. *Annual Review of Neuroscience*, 36, 337–359.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology*, 55, 349–374.

Vyas, S., Golub, M. D., Sussillo, D., e Shenoy, K. V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience*, 43, 249–275.

Apprendimento automatico e rappresentazioni geometriche

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., e Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4), 18–42.

Cohen, T. S., e Welling, M. (2016). Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning* (ICML 2016), 2990–2999.

Higgins, I., Amos, D., Pfau, D., Racanière, S., Rezende, D., e Mohamed, S. (2018). Towards a definition of disentangled representations. Preprint arXiv:1812.02230.

Nickel, M., e Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30* (NIPS 2017), 6338–6347.

Peng, H., Li, J., He, Y., Liu, Y., Bao, M., Wang, L., Song, Y., e Yang, Q. (2018). Large-scale hierarchical text classification with recursively regularized deep graph-CNN. In *Proceedings of the 27th International World Wide Web Conference* (WWW 2018), 1063–1072.

Sarkar, R. (2011). Low distortion Delaunay embedding of trees in hyperbolic plane. In *Proceedings of the 19th International Symposium on Graph Drawing* (GD 2011), Lecture Notes in Computer Science, vol. 7034, 355–366. Springer, Heidelberg.

Teoria dell'informazione integrata e coscienza

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, 4(1), niy007.

Oizumi, M., Albantakis, L., e Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLOS Computational Biology*, 10(5), e1003588.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3), 216–242.

Tononi, G., Boly, M., Massimini, M., e Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.

Psicofisica e spazio percettivo

Ayama, M., Nakatsue, T., e Kaiser, P. K. (1987). Constant hue loci of unique and binary hues at 10, 100, and 1000 Td. *Journal of the Optical Society of America A*, 4(7), 1136–1144.

Purdy, D. M. (1931). Spectral hue as a function of intensity. *American Journal of Psychology*, 43(4), 541–559.

Wyszecki, G., e Stiles, W. S. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2a ed. Wiley, New York.

Trasporto ottimale

Kantorovich, L. V. (1942). On the translocation of masses. *Doklady Akademii Nauk SSSR*, 37(7–8), 227–229. Trad. ingl. in *Journal of Mathematical Sciences*, 133(4), 1381–1382 (2006).

Villani, C. (2003). *Topics in Optimal Transportation*. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (RI).

Villani, C. (2009). *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, vol. 338. Springer, Berlin.

Teoria delle categorie applicata alle neuroscienze e alla cognizione

Fong, B., e Spivak, D. I. (2019). *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. Cambridge University Press, Cambridge.

Spivak, D. I. (2014). *Category Theory for the Sciences*. MIT Press, Cambridge (MA).